DOCUMENT RESUME

ED 168 901 SE 027 551

AUTHOR	Wright, David; And Others
TITLE	Three Assessments of Science, 1969-77: Technical Summary.
INSTITUTION	Education Commission of the States, Denver, Colo. National Assessment of Educational Progress.
SPONS AGENCY	National Center for Education Statistics (DHEW), Washington, D.C.
REPORT NO	NAEP-08-S-21
PUB DATE	Apr 79
CONTRACT	OEC-0-74-0506
NOTE	97p.; Contains occasional light and broken type; Certain tables contain small print
AVAILABLE FROM	National Assessment of Educational Progress, Suite 700, 1860 Lincoln St., Denver, Colorado 80295 ($10.60)
EDRS PRICE	MF01/PC04 Plus Postage.
DESCRIPTORS	*Achievement; Adults; *Age Groups; *Educational Assessment; Educational Research; Elementary Secondary Education; *National Surveys; *Science Education; Scientific Enterprise
IDENTIFIERS	National Assessment of Educational Progress

ABSTRACT

This technical summary describes the changes in science performance on exercises included in both the first and second science assessments and on exercises included in both the second and third science assessments conducted by the National Assessment of Educational Progress (NAEP). Using the same exercises for adjacent assessments, with some exercises common to all three, National Assessment was able to measure improvements and declines in achievement between 1969-70 and 1976-77. Each assessment utilized a deeply stratified, multistage probability sample design and a professional data collection staff. To the extent possible, administration conditions were kept constant across assessments. The document contains a brief introduction and four chapters: chapter 1 contains background information about the project, chapter 2 presents national results by age levels (9 years, 13 years, 17 years), chapter 3 reports group results for 9-, 13- and 17-year olds, and chapter 4 contains a discussion of the adult science assessments. Four appendices are included: (1) A-Technical Procedures: Sampling and Estimation of Standard Errors; (2) B-Estimated Population Proportions of Reporting Groups Based on National Assessment Samples, 1969-70, 1972-73, and 1975-76; (3) C-Changes in Procedures Between Assessments; and (4) D-Nonresponse in Assessment Samples. (PEB)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED168901

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

SCOPE OF INTEREST NOTICE

The ERIC Facility has assigned this document for processing to: SE TM

In our judgement, this document is also of interest to the clearinghouses noted to the right. Indexing should reflect their special points of view.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Theodore Pratt

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

THREE ASSESSMENTS OF SCIENCE, 1969-77: TECHNICAL SUMMARY

National Center for Education Statistics

NAEP

SE 027 551

This report is made pursuant to contract No. OEC-0-74-0506. The amount charged to the Department of Health, Education, and Welfare for the work resulting in this report (inclusive of the amounts so charged for any prior reports submitted under this contract) is $27,530,530. The names of the persons, employed or retained by the contractor, with managerial or professional responsibility for such work, or for the content of the report are as follows: Roy Forbes (see Acknowledgments).

The cost figure cited above represents the total amount of money expended since late 1973 on assessments in art, career and occupational development, reading, writing, social studies/citizenship, science, basic life skills, mathematics and consumerism, resulting, to date, in numerous reports, papers, articles, presentations and assessment materials, many of which are used in state and local assessment programs. A complete list of all such materials is available upon request.

THREE ASSESSMENTS OF SCIENCE, 1969-77:

Technical Summary

Report No. 08-S-21

by the
National Assessment of Educational Progress

Education Commission of the States
Suite 700, 1860 Lincoln Street
Denver, Colorado 80295

April 1979

Prepared under contract No. OEC-0-74-0506 with the National Center for Education Statistics, U.S. Department of Health, Education, and Welfare, Education Division. Contractors undertaking such projects are encouraged to express freely their professional judgment. This report, therefore, does not necessarily represent positions or policies of the Education Division, and no official endorsement should be inferred.

NATIONAL CENTER FOR EDUCATION STATISTICS

"The purpose of the Center shall be to collect and disseminate statistics and other data related to education in the United States and in other nations. The Center shall ... collect, collate, and, from time to time, report full and complete statistics on the conditions of education in the United States; conduct and publish reports on specialized analyses of the meaning and significance of such statistics; ... and review and report on education activities in foreign countries." -- Section 406(b) of the General Education Provisions Act, as amended (20 U.S.C. 1221e-1).

U.S. Department of Health, Education, and Welfare
Joseph A. Califano Jr., *Secretary*

Education Division
Mary F. Berry, *Assistant Secretary for Education*

National Center for Education Statistics
Marie D. Eldridge, *Administrator*

The National Assessment of Educational Progress is funded by the National Center for Education Statistics. It is under contract with the Education Commission of the States. It is the policy of the Education Commission of the States to take affirmative action to prevent discrimination in its policies, programs and employment practices.

TABLE OF CONTENTS

LIST OF TABLES AND FIGURES v

FOREWORD ix

ACKNOWLEDGMENTS xi

INTRODUCTION xiii
Organization of the Report xiv

CHAPTER 1 BACKGROUND 1
History of Objectives and Exercise Development 1
Sampling and Data Collection 3
Scoring 5
Measures of Achievement 6
Estimating Variability in Achievement Measures 8
Controlling Nonrandom Errors 9

CHAPTER 2 NATIONAL RESULTS 11
Results for 9-Year-Olds 11
Results for 13-Year-Olds 11
Results for 17-Year-Olds 14
Summary 14

CHAPTER 3 GROUP RESULTS FOR 9-, 13- AND 17-YEAR-OLDS 17
Definitions of Reporting Groups 17
Group Results 19

CHAPTER 4 THE ADULT SCIENCE ASSESSMENTS 27
The 1969 Assessment of Young Adults 27
The 1972-73 and 1977 Assessments of Young Adults 27
Differences Between the 1972-73 and 1977 Assessments 30
National Results 33
Group Results for Young Adults 33

APPENDIX A TECHNICAL PROCEDURES: SAMPLING AND ESTIMATION OF STANDARD ERRORS 37
Sampling 37
Estimation of Standard Errors 42
References 50

APPENDIX B ESTIMATED POPULATION PROPORTIONS OF REPORTING GROUPS BASED ON NATIONAL ASSESSMENT SAMPLES, 1969-70, 1972-73 AND 1975-76 51

APPENDIX C CHANGES IN PROCEDURES BETWEEN ASSESSMENTS 57
Definition of Variables 57
Data Collection Procedures. 60
"I Don't Know" Responses and Nonresponse. 67
Released and Reassessed Exercises 69

APPENDIX D NONRESPONSE IN ASSESSMENT SAMPLES 75

LIST OF TABLES AND FIGURES

TABLE 1-1. Numbers of Respondents for the Science Assessments, Ages 9, 13 and 17 4

TABLE 2-1. Mean Percentages of Correct Responses in Three Assessments and Changes in Percentages for All Exercises and Selected Exercise Classifications, Age 9. 12

TABLE 2-2. Mean Percentages of Correct Responses in Three Assessments and Changes in Percentages for All Exercises and Selected Exercise Classifications, Age 13 13

TABLE 2-3. Mean Percentages of Correct Responses in Three Assessments and Changes in Percentages for All Exercises and Selected Exercise Classifications, Age 17 15

TABLE 3-1. Reporting-Group Mean Differences in Percentage Correct From the Nation for 1970, 1973 and 1977; Changes in Mean Differences From 1970 to 1973 and From 1973 to 1977; and Standard Errors for the Total Change Exercise Sets at Age 9 21

TABLE 3-2. Reporting-Group Mean Differences in Percentage Correct From the Nation for 1969, 1972 and 1976; Changes in Mean Differences From 1969 to 1972 and From 1972 to 1976; and Standard Errors for the Total Change Exercise Sets at Age 13. 22

TABLE 3-3. Reporting-Group Mean Differences in Percentage Correct From the Nation for 1969, 1973 and 1977; Changes in Mean Differences From 1969 to 1973 and From 1973 to 1977; and Standard Errors for the Total Change Exercise Sets at Age 17. 23

FIGURE 3-1. Sex, Race, Region and Parental-Education Reporting-Group Mean Differences in Percentage Correct From the Nation in Three Assessments, Ages 9, 13 and 17 24

FIGURE 3-2. Type-of-Community, Size-of-Community and Grade-in-School Reporting-Group Mean Differences in Percentage Correct From the Nation in Three Assessments, Ages 9, 13 and 17 25

TABLE 4-1. Characteristics of the 1972-73 and 1977 Young Adult Assessments 29

TABLE 4-2. Screening and Response-Rate Data for the 1972-73 and 1977 Assessments of Young Adults. 31

TABLE 4-3. Mean Percentages Correct, Changes in Mean Percentages Correct and Standard Errors From 1972-73 to 1977 for Young Adults and for 17-Year-Olds on Selected Exercises 34

TABLE 4-4. Estimated Population Proportions, Mean Differences in Percentage Correct From the Nation for Young Adult Reporting Groups Between 1972-73 and 1977, Changes in Mean Differences and Standard Errors 35

TABLE A-1. Number of PSUs and Schools Within PSUs Selected in 1969-70, 1972-73 and 1976-77 39

TABLE A-2. Average Number of Respondents in Reporting Groups Taking a Package of Exercises, by Age and Assessment Year 47

TABLE A-3. Number of Change Exercises in Various Content Categories and Number of Packages in Which Those Exercises Appeared in 1969-70 and 1972-73, Ages 9, 13 and 17. 48

TABLE A-4. Number of Change Exercises in Various Content Categories and Number of Packages in Which Those Exercises Appeared in 1972-73 and 1976-77, Ages 9, 13 and 17. 49

TABLE B-1. Definition of Smoothing Cells for Adjusting Population Proportions. 53

TABLE B-2. Estimated Population Proportions of National Assessment Reporting Groups for Ages 9, 13 and 17 in 1969-70, 1972-73 and 1976-77. 54

TABLE B-3. Smoothing-Cell Proportions, Smoothed Proportions and Weight Adjustment Factors: Hypothetical Example for 13-Year-Old Males Taking Package 1 in 1976. 56

TABLE C-1. Mean Change in Percentages of Correct Responses Between 1972-73 and 1976-77 by Package Location in 1972-73 and 1976-77, Ages 9, 13 and 17. 63

FIGURE C-1. Mean Change in Percentage of Correct Response From 1972-73 to 1976-77 by Position in Package, Ages 9, 13 and 17 64

TABLE C-2. Mean Change in Percentages of Nonresponses Between 1972-73 and 1976-77 by Package Location in 1972-73 and 1976-77, Ages 9, 13 and 17. 65

FIGURE C-2. Mean Change in Percentage of Nonresponse From 1972-73 to 1976-77 by Position in Package, Ages 9, 13 and 17. 66

TABLE C-3. Mean Percentage Responding "I Don't Know" in 1969-70, 1972-73 and 1976-77 for Exercises Used to Measure Change From 1969-70 to 1972-73 and From 1972-73 to 1976-77, Ages 9, 13 and 17 68

TABLE C-4. Mean Percentage of Nonresponse in 1969-70, 1972-73 and 1976-77 for Exercises Used to Measure Change From 1969-70 to 1972-73 and from 1972-73 to 1976-77, Ages 9, 13 and 17 70

TABLE C-5. Mean Changes in Percentages of Correct Responses From 1970 to 1977 for Released and Unreleased Exercises by Content Classification, Age 9 71

TABLE C-6. Mean Changes in Percentages of Correct Responses From 1969 to 1976 for Released and Unreleased Exercises by Content Classification, Age 13. 72

TABLE C-7. Mean Changes in Percentages of Correct Responses From 1969 to 1977 for Released and Unreleased Exercises by Content Classification, Age 17. 73

TABLE D-1. Number of Students Assessed and Percent of Sample Covered by Age, Assessment Year and Type of Administration. 77

FOREWORD

When the U.S. Office of Education was chartered in 1867, one charge to its commissioners was to determine the nation's progress in education. The National Assessment of Educational Progress (NAEP) was initiated a century later to address, in a systematic way, that charge.

Each year since 1969, National Assessment has gathered information about levels of educational achievement across the country and reported its findings to the nation. NAEP surveys the educational attainments of 9-year-olds, 13-year-olds, 17-year-olds and young adults, ages 26-35, in 10 learning areas: art, career and occupational development, citizenship, literature, mathematics, music, reading, science, social studies and writing. Different learning areas are assessed every year, and all areas are periodically reassessed in order to measure change in educational achievement. National Assessment has interviewed and tested more than 720,000 young Americans since 1969.

Learning-area assessments evolve from a consensus process. Each assessment is the product of several years of work by a great many educators, scholars and lay persons from all over the nation. Initially, these people design objectives for each subject area, proposing general goals they feel Americans should be achieving in the course of their education. After careful reviews, these objectives are given to exercise (item) writers, whose task it is to create measurement tools appropriate to the objectives.

When the exercises have passed extensive reviews by subject-matter specialists, measurement experts and lay persons, they are administered to probability samples. The people in these samples are chosen in such a way that they represent the national population. Therefore, on the basis of the performance of about 2,500 9-year-olds on a given exercise, we can make generalizations about the probable achievement of all 9-year-olds in the nation. Performance is reported in terms of the percentages of young people correctly answering a given exercise or set of exercises; changes in performance are the differences between the percentages of young people correctly answering a given exercise or set of exercises from one point in time to another.

After assessment data have been collected, scored and analyzed, National Assessment publishes reports to disseminate the results as widely as possible. Not all exercises are released for publication. Because NAEP will readminister some of the same exercises in the future to determine whether the performance level of Americans has increased, decreased or remained the same, it is essential that they not be released in order to preserve the integrity of the study.

See the inside back cover of this report for a complete listing of additional reports on science assessments.

ACKNOWLEDGMENTS

Assessing science achievement of young Americans throughout the nation is an undertaking of major proportions. Certainly it could not have become a reality without substantial contributions by a great number of people, not the least of whom are the students, teachers and administrators who cooperated so generously.

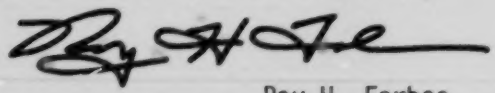
Dozens of consultants -- both subject-matter specialists and lay persons -- developed and reviewed these materials under the general guidance of the National Assessment of Educational Progress (NAEP) staff. Administration of exercises was handled by the Research Triangle Institute and the Measurement Research Center (MRC). Scoring was carried out by MRC and by NAEP staff members.

Special acknowledgment must go to the Analysis Advisory Committee (ANAC) of National Assessment, who developed the details of the plan for measuring change. Without their consistent support and advice, on both technical and general matters, this work would not have been achieved.

Finally, the collaborative effort of the entire National Assessment staff is gratefully acknowledged. Thanks are due to those persons in the Development, Analysis and Research Department who were responsible for monitoring and implementing development, scoring, analysis, technical documentation and quality-control procedures; those persons in the Operations Department who were responsible for preparing the packages and monitoring field procedures; those persons in the Data Processing Department who were responsible for the programming necessary to complete the analyses; and those persons in the Publications Department who were responsible for reporting data and editing documents related to these assessments.

Major technical contributions to analysis planning and implementation were made by Nancy Burton, Eugene Johnson, Robert Larson and Donald Searls. David Wright coordinated the development and implementation of scoring and analysis procedures.

The production support necessary for the actual completion of this report was provided by JoAnn Esslinger and Fred Daniels of the Development, Analysis and Research Department; and Marci Reser and Paula Pitchford of the Publications Department. David Wright had major responsibility for report preparation, with Robert Larson, Ina Mullis and Barbara Ward making significant contributions.



Roy H. Forbes
Director

INTRODUCTION

The National Assessment of Educational Progress has completed three assessments of science. They were conducted in 1969-70, 1972-73 and 1976-77.[1]

This report summarizes changes in science performance on exercises included in both the first and second science assessments and on exercises included in both the second and third science assessments. Using the same exercises for adjacent assessments, with some exercises common to all three, National Assessment was able to measure improvements and declines in achievement between 1969-70 and 1976-77.

Each assessment utilized a deeply stratified, multistage probability sample design and a professional data collection staff. To the extent possible, administration conditions were kept constant across assessments. Appendix C documents the procedural changes that have occurred between the first and third assessments.

Changes in science performance in this report have been summarized according to the 1972-73 science objectives[2] and by type of science (content). The content clusters comprise biology, physical science and other, or unclassified. An additional summary has been included for exercises that were administered in all three assessments of 9-, 13- or 17-year-olds enrolled in school.

National Assessment has published a number of reports related to science. A complete list is included on the inside back cover of this report. Reports most relevant to this technical summary include:

- Report 1 -- Science: National Results (July 1970). Contains released exercises from the first science assessment and technical documentation of methodology.

---

[1]The assessment schedule varied for each age level. The actual administration dates were:

Age 9:	January through February 1970, 1973 and 1977
Age 13:	October through December 1969, 1972 and 1976
Age 17:	March through May 1969, 1973 and 1977
Young Adults:	October 1972 through May 1973 and May through July 1977

[2]Science Objectives for 1972-73 Assessment (Denver, Colo.: Education Commission of the States, National Assessment of Educational Progress, 1972).

- Report 04-S-20 -- Changes in Science Performance, 1969-73: Exercise Volume (December 1975). Contains change exercises that were released after the 1972-73 science assessment with national results for all responses and correct response results for sex, race and region.

- Report 04-S-20 -- Changes in Science Performance, 1969-73: Exercise Volume, Appendix (two volumes, April 1977). Contains all exercises released after the 1972-73 science assessment, with percentages and standard errors as well as change statistics for region, sex, race, parental education and size and type of community.

- Report 04-S-21 -- Science Technical Report: Summary Volume (May 1977). Contains detailed methodological documentation of the 1969-70 and 1972-73 science assessments as well as summary data for objectives and content classifications.

- Report 03/04-GIY -- General Information Yearbook (December 1974). Contains a condensed description of National Assessment methodology with emphasis on the 1971-72 and 1972-73 assessments.

- Report 08-S-00 -- Three National Assessments of Science: Changes in Achievement, 1969-77 (June 1978). Contains a capsule description of changes in science achievement between 1969 and 1977 with interpretive comments by a group of science educators.

- The Third Assessment of Science, 1976-77: Released Exercise Set (May 1978). Contains exercises released after the 1976-77 science assessment, including exercises used to measure changes in achievement from 1969-70 and 1972-73.

- Technical Appendix to the Third Assessment of Science, 1976-77: Released Exercise Set (December 1978). Contains 1976-77 percentages of correct responses and standard errors for correct responses to all released cognitive exercises. Variables include race, sex, region, community size and grade.

Organization of the Report

The first chapter of this report presents a history of the development of the science objectives and exercises and describes procedures for sampling, data collection, scoring and analysis.

The second chapter summarizes changes in mean percentages of acceptable responses for each in-school age group. Summaries are presented for all exercises as well as the 1972-73 science objectives, content categories and exercises administered in all three science assessments.

The third chapter describes changes in performance for various school-age subpopulations: geographic region, sex, race, level of parental education, type of community, size of community and grade in school.

The fourth chapter describes changes in performance for young adults, ages 26-35, between 1973 and 1977.

CHAPTER 1

BACKGROUND

History of Objectives and Exercise Development

The exercises used to report changes in science achievement measure broad education objectives, which represent a consensus of educators, subject-matter experts and interested lay persons about what young Americans should know and be able to do. These objectives are not an attempt to mandate behavior and value systems; rather, they represent goals that a diverse group of people identified as desirable for young Americans to accomplish.

Objectives for the 1969-70 science assessment were developed by the Educational Testing Service in 1965.[1] During 1969 through 1971, the objectives were reorganized for the 1972-73 assessment.[2] The major 1972-73 objectives were:

1. Know the fundamental aspects of science.

2. Understand and apply the fundamental aspects of science in a wide range of problem situations.

3. Appreciate the knowledge and processes of science, the consequences and limitations of science, and the personal and social relevance of science and technology in our society.

Subobjectives for each objective consisted of Fundamental Aspects of Science and the Scientific Enterprise. Fundamental aspects were further subdivided into: facts and simple concepts, laws (principles), conceptual schemes and inquiry skills.

The number of exercises used to measure change between assessments by age group and objective is shown on the following page.

---

[1] Science Objectives, 1969-70 Assessment (Ann Arbor, Mich.: Committee on Assessing the Progress of Education, 1969), available through the National Assessment offices.

[2] Science Objectives for 1972-73 Assessment (Denver, Colo.: Education Commission of the States, National Assessment of Educational Progress, 1972).

Assessments	Age	Know	Understand/ Apply	Appreciate	Total Number of Exercises
1969-70 to 1972-73	9	40	47	5	92
	13	37	28	2	67
	17	34	30	0	64
1972-73 to 1976-77	9	32	36	3	71
	13	38	37	0	75
	17	31	37	2	70

The process of developing objectives and exercises to assess performance in a subject area across time is a difficult task. There must be a sufficient number of identical items to measure change reliably; on the other hand, the assessment must keep current with changing curriculum objectives. Therefore, after each assessment some items are released to the public and some are kept secure for the purpose of measuring change. Before the next assessment, the objectives are reviewed and revised, and new items are written to measure the revised objectives.

For the 1976-77 assessment, a somewhat different approach to objectives development was taken. Science consultants and National Assessment staff agreed that the 1972-73 objectives represented an excellent statement of the purposes and goals of science education but were not specific enough to provide a clear guide for writing assessment exercises. For assessment purposes, a two-dimensional grid was defined.[3] The first dimension is similar to the 1972-73 objectives, with four levels: knowledge; comprehension; application; and analysis, synthesis and evaluation. The second dimension divides the domain of science into three major areas: content, the body of science knowledge; the process by which the body of knowledge comes about; and science and society, the implications of that body of knowledge for mankind. Each of these is further subdivided into specific components. Within each cell of the grid, specific objectives were developed to guide item development.

While the 1976-77 objectives have not been used as a basis with which to summarize the changes in achievement from preceding assessments, they have been used to summarize cognitive achievenent in the 1976-77 assessment[4] and will be used to summarize change measures from 1976-77 to the next assessment of science.

---

[3]Science Objectives for the 1976-77 Assessment (Denver, Colo.: Education Commission of the States, National Assessment of Educational Progress, forthcoming).

[4]Science Achievement in the Schools, Report 08-S-01, 1976-77 Assessment (Denver, Colo.: Education Commission of the States, National Assessment of Educational Progress, 1978).

Many people from across the country have been involved in the development of objectives and items for these assessments. Subject-matter specialists, measurement experts and lay persons not only helped develop the objectives, they also participated in reviewing and revising exercises. All newly developed items were field-tested with students representative of high- and low-performing groups. Before and after each "tryout" assessment, the exercises were discussed by panels of reviewers, many of whom represented minority groups, to guard against the possibility of racial, ethnic or sexual bias.

Sampling and Data Collection

Each year National Assessment selects respondents at ages 9, 13 and 17 using a deeply stratified, multistage probability sample design.[5] This sample design guarantees that each respondent is selected with a known probability; hence, each respondent represents a known fraction of the entire population at that age level. By weighting each respondent's performance inversely to his or her probability of selection, National Assessment can make appropriate generalizations about the entire population of 9-year-olds, 13-year-olds and 17-year-olds enrolled in school.

National Assessment does not follow up specific individuals from one assessment to the next. In other words, the students who participated in the 1969-70 or the 1972-73 assessments are not the same ones who participated in 1976-77. However, in each assessment year, participants are carefully selected to represent each age level. For example, National Assessment assessed one probability sample of 9-year-olds to ascertain science achievement in 1970 and totally different probability samples of 9-year-olds in 1973 and 1977. Each was a sample of the population of students who were 9 years old during that assessment year. Thus, when we say that 9-year-olds' achievement declined between 1970 and 1973, we mean that students who were 9 years old in 1970 correctly answered the same questions more often than those who were 9 years old in 1973.

The three school-age populations selected for each of the science assessments were defined as follows:

---

[5]See Appendix A for technical details about National Assessment sampling procedures.

Age	1969-70 Assessment	1972-73 Assessment	1976-77 Assessment
9	Born in 1960	Born in 1963	Born in 1967
13	Born in 1956	Born in 1959	Born in 1963
17	Born between October 1951 and September 1952	Born between October 1955 and September 1956	Born between October 1959 and September 1960

The populations were further restricted to students enrolled in public or private schools who were neither in institutions nor too functionally handicapped to respond to assessment exercises.[6]

Once the exercises were selected for the assessments, they were assembled into booklets that were administered to probability samples of each appropriate age group. Not all students responded to all exercises. Each booklet or group of exercises was administered to a representative sample of about 2,500 9-, 13- or 17-year-olds. The approximate numbers of respondents who participated in the science assessments are shown in Table 1-1.

TABLE 1-1. Numbers of Respondents for the Science Assessments, Ages 9, 13 and 17

Age	1969-70	1972-73	1976-77
9	19,468	20,862	17,345
13	21,696	23,507	25,653
17	22,913	25,865	29,140

In order for an assessment to measure changes in performance reliably, it must replicate testing conditions as nearly as possible. Thus, items used to measure change are as nearly identical in wording and format in each assessment as is possible. National Assessment further attempts to keep administration procedures constant by tape-recording instructions and items and by using trained administrators, rather than classroom personnel, to conduct assessments. A discussion of changes that have taken place over the course of the three assessments can be found in Appendixes A and C.

---

[6]The 1969 and 1973 assessments of 17-year-olds included samples of dropouts and early graduates. Funding limitations precluded a similar sample in 1977. Thus, results in this report are limited to 17-year-olds enrolled in school.

Scoring

It is also essential that identical scoring procedures be used in each assessment if data are to be used to measure change. Both multiple-choice and open-ended exercises were included in the science assessments. Not more than six open-ended exercises per age were included in change summaries for 1972-73 to 1976-77. One open-ended exercise for 17-year-olds was included in 1969-70 to 1972-73 summaries. Individually administered experiments were included in both the 1969-70 and 1972-73 assessments. Because of technical difficulties with apparatus and scoring protocols, none of the experiments were included in change summaries for the first two assessments. Funding limitations precluded the use of individually administered experiments in the 1976-77 assessment.

Responses to multiple-choice items were marked directly in the assessment booklets. The booklets were optically scanned and edited by both computer and scoring staffs to ensure reliable scoring.

Only about three to five open-ended exercises per age group were available for measuring change between the 1969-70 and 1972-73 assessments. One exercise was rescored for age 17 and included in change summaries. The remainder were omitted from summaries because of the questionable comparability of scoring procedures.

Scoring comparability for open-ended items was achieved between 1972-73 and 1976-77 by rescoring the 1972-73 responses simultaneously with the scoring of 1976-77 responses. Four highly trained scorers with previous assessment scoring experience coded the responses for each age group as assessment booklets were received from the data collection staff.

Scoring for each age group took 8 to 12 weeks. At the beginning of scoring for each age group, the scorers were trained by the Measurement Research Center scoring director, a science consultant and the National Assessment science analyst. The scoring guide for each exercise was presented and discussed. Sample responses from both the 1972-73 and 1976-77 assessments were independently coded by both scorers and trainers and scores were compared for consistency. Scoring guides were clarified and revised, if necessary, and more sample responses were scored until near-perfect consistency was achieved.

To help maintain quality control and identify problems, 10% of each scorer's work was independently scored by another, usually within one or two weeks of each other. Agreement between scorers, on about 250 to 260 responses per exercise, ranged from 96 to 100% on the open-ended exercises included in change summaries, as shown below.

Age	Number of Exercises	Range of Percent of Agreement on 10% Subsample
9	5	96 to 100%
13	6	96 to 99%
17	6	97 to 100%

These figures indicate how consistently a small group of highly trained scorers can score the same set of papers.

Measures of Achievement

The basic measure of achievement reported by National Assessment is the percentage responding acceptably to a given item. This percentage is an estimate of the percentage of 9-, 13- or 17-year-olds who would respond acceptably to a given item if every 9-, 13- or 17-year-old in the country were assessed.

Percentages of correct responses are used because each item is designed as a separate measure of some aspect of an objective or subobjective. The purpose of National Assessment is to discover if more or fewer people are able to answer these items correctly -- and thus meet the objectives -- over the years.

Procedures for estimating percentages of acceptable responses to exercises are dependent on the sample design. Each response by an individual is weighted and multiplied by an adjustment factor for nonresponse.[7] An estimate of the percentage of a particular age group that would have responded to an exercise acceptably if the entire age group were assessed is defined as the weighted number of acceptable responses divided by the weighted number of all responses. A similar ratio of weights is used to estimate percentages of acceptable responses for reporting groups or subpopulations of interest.[8]

The difference between the percentage of acceptable responses for a reporting group and that of the entire age group on an exercise describes the performance of any reporting group relative to the entire age group. This difference is a positive number if the group achieves a higher percentage than the entire age group and is a negative number if the group achieves a lower percentage. For example, a group performance of +1.8 indicates that the percentage of acceptable responses for the group is 1.8 percentage points higher than the national percentage of acceptable responses for a particular age level.

Increases or decreases in the percentage of acceptable responses between two assessments are estimated by finding the difference between percentages obtained from each assessment. A positive difference indicates an increase,

---

[7]Appendix D discusses nonresponse in assessment samples.

[8]Following the 1976-77 assessment, a weighting-class adjustment procedure was used to dampen fluctuations in estimated population proportions across the eight assessments conducted between 1969-70 and 1976-77. Documentation of this procedure and estimated population proportions are included in Appendix B. Consequently, the estimated percentage of correct responses in this report and Three National Assessments of Science: Changes in Achievement, 1969-77, Report 08-S-00, may deviate slightly from the figures in earlier science change reports.

and a negative difference indicates a decrease in the percentage of students who responded acceptably from one assessment to the next. These differences, or change measures, are used to indicate trends in achievement, or performance, for an age level or subpopulation of interest. Changes in group differences from the national performance between two assessments are used to indicate the relative trend of a group compared to the national trend of the age group.

To present a general picture of changes in achievement, National Assessment summarizes the gains or losses on each exercise (either for the entire learning area or for some integral set of exercises) by using the mean, or arithmetic average, of the changes in percentages of acceptable responses to the exercises. During the first years of the assessment, the median was used as the principal summary measure. However, the mean was chosen as the principal summary measure of change after extensive investigation showed empirically that it was more suitable for National Assessment change data than alternative measures.[9] In addition, the mean is an easily understood and fairly well-known statistic and has simple arithmetic properties useful for the analysis of differences or change measures -- in particular, the difference between means is the same as the mean difference. This property allows us to describe accurately the mean change as the difference between mean percentages of acceptable responses from one assessment to the next. Mean percentages for the science assessments are used throughout this report to simplify descriptions of change. Its use does not signify that the mean is the best summary statistic to use in each assessment separately, nor do we intend that the mean percentage should be construed as an average test score.

Unless the items summarized in the mean percentages of acceptable responses are identical, the means of one age group should not be compared to the means of another, since their values reflect the choice of exercises in addition to the performance of the students. When only a few exercises are summarized by a mean, we should be especially cautious in interpreting results, since a small set of exercises might not adequately cover the wide range of potential behaviors included under a given objective or subobjective. The mean should be interpreted literally as the arithmetic average of the percentage of acceptable responses obtained from National Assessment samples on a specific set of exercises.

---

[9]Twenty-two empirical distributions of change measures from the 1969-70 and 1972-73 science assessments were used to generate Monte Carlo simulations of sampling distributions for several measures of central location. In addition to the mean and median, other measures of central location that were considered in the simulation studies included the average of the extremes, two forms of biweighted estimates and three forms of weight-matching estimators described by John W. Tukey in the research report, "Some Considerations on Locators Apt for Some Squeezed-Tail (and Stretched-Tail) Parents" (paper prepared in connection with research at Princeton University supported by the Army Research Office, summer 1975). In almost every case, the sampling stability of the mean change was as good as or better than that of the other measures studied.

In the analysis of National Assessment's achievement measures, notice that the differences in performance among assessment years, among groups and among ages are most useful. By maintaining the same item or set of items in making these comparisons, we have a reasonable indicator of whether more or fewer people know or can do something judged important.

Estimating Variability in Achievement Measures

National Assessment uses a national probability sample at each age level to estimate the proportion of people who would successfully complete an exercise. The particular sample selected is one of a large number of all possible samples of the same size that could have been selected with the same sample design. Since an achievement measure computed from each of the possible samples would differ from one sample to another, the standard error of this statistic is used as a measure of the sampling variability among achievement measures from all possible samples. A standard error, based on one particular sample, serves to estimate that sampling variability.

In the interest of sampling and cost efficiencies, National Assessment uses a complex, stratified, multistage probability sample design. Typically, complex designs do not provide for unbiased or simple computation of sampling errors. A reasonably good approximation of standard-error estimates of acceptable response percentages is obtained by applying the jackknife procedure[10] to first-stage sampling units within strata. Standard errors for achievement measures such as group differences, mean percentages or mean group differences for a particular assessment year are estimated directly, taking advantage of features of the jackknife procedure that are generic to all of these statistics.[11] Since samples for different assessments are independent, the standard errors of the differences in achievement measures between assessments can be estimated simply by the square root of the sum of squared standard errors from each of the assessments.

The standard error provides an estimate of sampling reliability for the achievement measures used by National Assessment. It is comprised of sampling error and other random error associated with the assessment of a specific item or set of items. Random error includes all possible nonsystematic error associated with administering specific exercises to specific students in spe-

---

[10]R.G. Miller Jr., "A Trustworthy Jackknife," Annals of Mathematical Statistics, No. 35 (1964), pp. 1594-1705; R.G. Miller Jr., "Jackknifing Variances," Annals of Mathematical Statistics, No. 39 (1968), pp. 567-82; F. Mosteller and J.W. Tukey, "Data Analysis Including Statistics," in Handbook of Social Psychology (2nd ed.), eds. E. Aronson and G. Lindzey (Reading, Mass: Addison-Wesley, 1968).

[11]See Appendix A for a more detailed description of National Assessment's computation of standard errors.

cific situations. Random differences among scorers for open-ended items are also included in the standard errors.

In this report, we designate with an asterisk item differences or mean differences that are at least twice as large as their standard errors. By so designating these differences, we are adopting the usual convention that differences this large would occur by chance in fewer than 5% of all possible replications of our sampling and data collection procedures.

Controlling Nonrandom Errors

Systematic errors can be introduced at any stage of an assessment -- exercise development, preparation of exercise booklets, design of administration procedures, field administration, scoring or analysis. These nonsampling, nonrandom errors rarely can be quantified, nor can the magnitude of the bias they introduce into our estimates be evaluated directly.

Systematic errors can be controlled in large part by employing uniform administration and scoring procedures and by requiring rigorous quality control in all phases of an assessment. If the systematic errors are the same from age to age or group to group, then the differences in percentages or mean percentages are measured with reduced bias because subtraction tends to cancel the effect of the systematic errors.

Similarly, the effect of systematic errors in different assessment years can be controlled by carefully replicating in the second assessment the procedures carried out in the first. Differences in achievement across assessment years will also be measures with reduced bias since subtraction will again tend to cancel systematic errors.

However, it is not possible for every condition or procedure to remain the same between assessments conducted several years apart. Improvements in field procedures and sample design have been made, school cooperation rates have improved slightly since the early assessments, packaging of exercises was not identical in each assessment, and shifts in the composition of categories of respondents have occurred over the years.[12]

---

[12]Appendix C examines some of these changes and discusses the possible effects of these systematic errors on the results in this report.

CHAPTER 2

NATIONAL RESULTS

This chapter presents national data on changes in science performance for 9-, 13- and 17-year-olds. Results are summarized for all exercises used to measure change from 1969-70 to 1972-73 and from 1972-73 to 1976-77. They are also summarized by the 1972-73 objectives and type of science, as well as the sets of exercises common to all three assessments.

Discussion of results is minimized since National Assessment has published a major descriptive report based on these data.[1] This chapter contains the results presented in that report, plus supplementary information.

Results for 9-Year-Olds

Table 2-1 contains the number of exercises, means and standard errors for each set of change exercises. Between 1970 and 1973, science achievement of 9-year-olds declined on most of the summary measures. The decline was not significant for biology or unclassified exercises, and the five exercises dealing with the objective of appreciation showed a significant increase. There was no overall change between 1973 and 1977. Achievement on physical science exercises declined significantly, but it increased significantly on unclassified exercises while achievement on biology exercises increased by almost two standard errors. The three appreciation exercises showed a significant increase in the percentage of correct responses.

Results for 13-Year-Olds

Table 2-2 contains the number of exercises and summary results for each set of change exercises. Results for 13-year-olds from 1969-72 were similar to those for 9-year-olds during the same time period. There was a significant overall decline in achievement. The decline on biology exercises was not significant, and performance on the unclassified science exercises showed no change, while there was a significant increase on the two exercises dealing with the appreciation objective. Between 1972 and 1976 there was no overall

---

[1] Three National Assessments of Science: Changes in Achievement, 1969-77, Report 08-S-00 (Denver, Colo.: Education Commission of the States, National Assessment of Educational Progress, 1978).

TABLE 2-1. Mean Percentages of Correct Responses in Three Assessments and Changes in Percentages for All Exercises and Selected Exercise Classifications, Age 9

	Mean % Correct				Mean % Correct			
	Number of Exercises	1970	1973	Change	Number of Exercises	1973	1977	Change
All exercises	92	61.0	59.8	-1.2*	71	52.3	52.2	- .1
Standard error		(.4)	(.4)	(.6)		(.4)	(.4)	(.6)
Type of science								
Biology	27	70.4	69.3	-1.0	24	57.8	59.2	1.4
Standard error		(.4)	(.4)	(.6)		(.4)	(.6)	(.7)
Physical science	50	56.7	55.2	-1.5*	42	47.5	46.2	-1.3*
Standard error		(.4)	(.5)	(.6)		(.4)	(.4)	(.6)
Unclassified	15	58.8	58.6	- .2	5	66.3	69.1	2.8*
Standard error		(.5)	(.6)	(.8)		(.7)	(.8)	(1.1)
1972-73 objective								
Know	40	66.3	64.5	-1.9*	32	54.8	54.2	- .6
Standard error		(.3)	(.6)	(.6)		(.5)	(.6)	(.7)
Understand and apply	47	55.0	54.0	-1.0*	36	47.7	47.8	- .2
Standard error		(.4)	(.5)	(.6)		(.4)	(.4)	(.6)
Appreciate	5	74.7	77.2	-2.5*	3	82.0	84.9	3.0*
Standard error		(.6)	(.7)	(.9)		(.7)	(.9)	(1.1)
Exercises used in all three assessments	30	64.8	63.7	-1.2*	30	63.7	62.9	- .8
Standard error		(.4)	(.4)	(.6)		(.4)	(.5)	(.6)

**Denotes differences greater than or equal to two standard errors.*

Note: Computations were performed prior to rounding to one decimal place.

TABLE 2-2. Mean Percentages of Correct Responses in Three Assessments and Changes in Percentages for All Exercises and Selected Exercise Classifications, Age 13

	Number of Exercises	Mean % Correct 1969	1972	Change	Number of Exercises	Mean % Correct 1972	1976	Change
All exercises	67	60.2	58.5	-1.7*	75	54.5	53.8	- .7
Standard errors		(.4)	(.5)	(.6)		(.4)	(.4)	(.6)
Type of science								
Biology	23	60.9	59.6	-1.3	23	61.1	62.0	.9
Standard error		(.8)	(.5)	(.7)		(.4)	(.5)	(.7)
Physical science	36	59.7	57.1	-2.6*	47	50.4	49.6	- .8
Standard error		(.4)	(.5)	(.7)		(.4)	(.4)	(.6)
Unclassified	8	64.7	65.4	.7	5	62.1	55.8	-6.3*
Standard error		(.6)	(.6)	(.8)		(.9)	(1.0)	(1.3)
1972-73 objective								
Know	37	60.0	58.5	-1.6*	38	56.5	55.4	-1.1
Standard error		(.3)	(.5)	(.6)		(.4)	(.5)	(.7)
Understand and apply	28	59.9	57.6	-2.2*	37	52.4	52.2	- .2
Standard error		(.4)	(.5)	(.7)		(.4)	(.4)	(.6)
Appreciate	2	66.0	69.8	3.9*	0	--	--	--
Standard error		(.7)	(.9)	(1.1)		--	--	--
Exercises used in all three assessments	13	63.3	61.4	-1.9*	23	61.4	59.7	-1.7*
Standard error		(.4)	(.5)	(.6)		(.5)	(.5)	(.7)

**Denotes differences greater than or equal to two standard errors.*

Note: Computations were performed prior to rounding to one decimal place.

change, although performance on five unclassified science exercises declined, as it did on the exercises carried over from the 1969 assessment.

Results for 17-Year-Olds

Table 2-3 contains summary results for 17-year-olds on each set of change exercises. Average achievement on all exercises declined between 1969 and 1973. These results are reflected in the other summaries for that time period: all three types of science, the two objectives for which exercises were available and the exercises common to all three assessments. Results between 1973 and 1977 were similar to those of the earlier time period. Achievement declined on all classifications of exercises, but declines were not significant for biology, the six unclassified science exercises or the two exercises measuring the appreciation objective.

Summary

Average science achievement of 9-, 13- and 17-year-olds declined between 1969-70 and 1972-73 on the overall summaries and most subclassifications. That trend continued for 17-year-olds but not for 9- and 13-year-olds between 1972-73 and 1976-77. Achievement on biology exercises appears to have stabilized, while the decline in achievement on physical science exercises might be slowing for 9- and 13-year-olds. A larger, more comprehensive set of science exercises will be available for measuring changes in performance between 1976-77 and the next assessment. Thus, data from the larger, fourth assessment, combined with results from the first three, might clarify trends.

TABLE 2-3. Mean Percentages of Correct Responses in Three Assessments and Changes in Percentages for All Exercises and Selected Exercise Classifications, Age 17

	Mean % Correct				Mean % Correct			
	Number of Exercises	1969	1973	Change	Number of Exercises	1973	1977	Change
All exercises	64	45.2	42.5	-2.8*	70	48.4	46.5	-1.9*
Standard error		(.3)	(.3)	(.5)		(.4)	(.4)	(.6)
Type of science								
Biology	20	52.3	51.1	-1.2*	19	53.3	52.2	-1.1
Standard error		(.4)	(.4)	(.6)		(.5)	(.5)	(.7)
Physical science	39	42.9	39.3	-3.5*	45	46.8	44.4	-2.4*
Standard error		(.4)	(.4)	(.5)		(.4)	(.4)	(.6)
Unclassified	5	35.6	32.1	-3.5*	6	44.8	43.8	-1.0
Standard error		(.6)	(.6)	(.8)		(.6)	(.7)	(1.0)
1972-73 objectives								
Know	34	49.9	47.0	-2.9*	31	50.5	49.3	-1.2*
Standard error		(.4)	(.4)	(.5)		(.4)	(.4)	(.6)
Understand and apply	30	40.0	37.3	2.7*	37	45.7	43.3	-2.5*
Standard error		(.4)	(.4)	(.6)		(.4)	(.5)	(.6)
Appreciate	0	--	--	--	2	65.4	63.2	-2.2
Standard error		--	--	--		(1.1)	(1.3)	(1.6)
Exercises used in all three assessments	23	44.6	42.3	-2.3*	23	42.3	39.9	-2.4*
Standard error		(.4)	(.4)	(.6)		(.4)	(.4)	(.6)

**Denotes differences greater than or equal to two standard errors.*

Note: Computations were performed prior to rounding to one decimal place.

CHAPTER 3

GROUP RESULTS FOR 9-, 13- AND 17-YEAR-OLDS

This chapter contains definitions of National Assessment reporting groups and summary results for the full sets of exercises used to measure change from 1969-70 to 1972-73 and from 1972-73 to 1976-77. Respondents were classified by their sex, race, region, highest level of parental education, type of community, size of community and grade levels. Estimated proportions for each subpopulation are listed in Appendix C.

Definitions of Reporting Groups

The definitions of the categories used in this report for 9-, 13- and 17-year-olds are given below.

Sex

Results are presented for males and females.

Race

Results are presented for blacks and whites.

Region

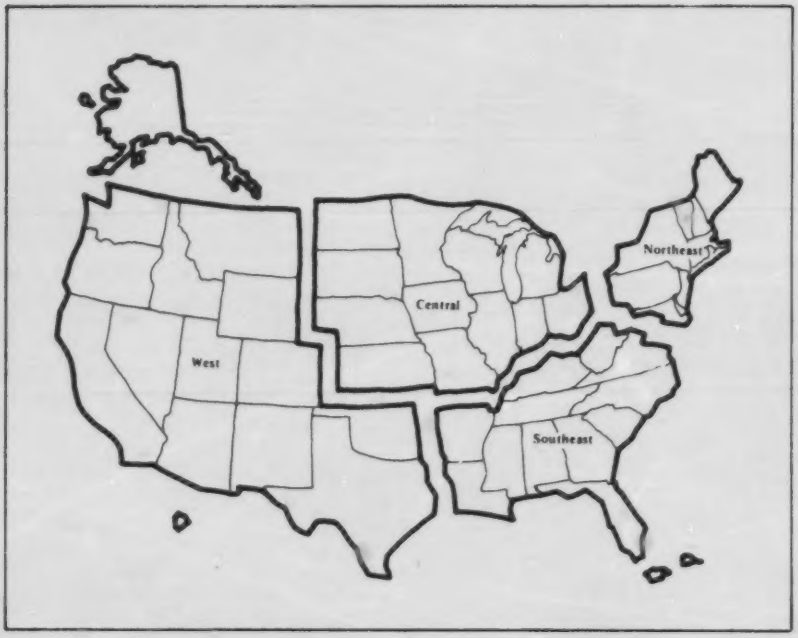
The country has been divided into four regions: Northeast, Southeast, Central and West. States included in each region are shown on the following page (see map).

Level of Parental Education

Three categories of parental-education levels are defined by National Assessment, based on students' reports about them. These categories are: those whose parents did not graduate from high school; those who have at least one parent who graduated from high school; and those who have at least one parent who has had some post high school education.[1]

---

[1]The form of the parental-education question was changed slightly after the 1969-70 assessment. Details are given in Appendix C.



Grade Level

Results are categorized for 9-year-olds in the 3rd or 4th grade, 13-year-olds in the 7th or 8th grade and 17-year-olds in the 10th, 11th or 12th grade.

Size of Community

Big city. Students in this group attend schools within the city limits of cities having a 1970 census population over 200,000.

Fringes around big cities. Students in this group attend schools within metropolitan areas (1970 U.S. Bureau of the Census urbanized areas) served by cities having a population greater than 200,000 but outside the city limits.

Medium city. Students in this group attend schools in cities having a population between 25,000 and 200,000, not classified in the fringes-around-big-cities category.

Smaller places. Students in this group attend schools in communities having a population less than 25,000, not classified in the fringes-around-big-cities category.

Type of Community

Communities in this category are defined by an occupational profile of the

area served by a school as well as by the size of the community in which the school is located.

Advantaged-urban (high-metropolitan) communities. Students in this group attend schools in or around cities with a population greater than 200,000 where a high proportion of the residents are in professional or managerial positions.

Disadvantaged-urban (low-metropolitan) communities. Students in this group attend schools in or around cities with a population greater than 200,000 where a high proportion of the residents are on welfare or are not regularly employed.

Extreme-rural communities. Students in this group attend schools in areas with a population under 10,000 where most of the residents are farmers or farm workers.

Group Results

The distinction between cross-sectional and longitudinal survey research designs is especially important to note in order to interpret changes in results for groups of respondents. National Assessment does not report changes for the same individuals; rather, it reports changes on the same types of groups of respondents, such as those living in the Southeast or those attending schools in rural areas. Thus, a group of respondents in one assessment might have a composition of people different from the same group defined in the same way in another assessment.

The longer the time between assessments, the more these groups might differ. The Southeast, for example, might become more urbanized or its racial composition might change because of migration between regions. The extreme-rural respondents in any given year are defined as the 10% in our sample attending the most rural schools; schools classified as extreme rural one year might not be the most rural in the next assessment because of population shifts, consolidation of schools, and so on. Every attempt has been made to keep the category definitions constant; however, we know some changes in the composition of these categories occurred between 1969-70 and 1976-77.[2]

Group results are computed by estimating the mean percentage correct for a reporting group in the same manner as described previously for the national

---

[2]U.S. Bureau of the Census, "Mobility of the Population of the United States: March 1970 to March 1973," Current Population Reports, Series P-20, No. 262 (Washington, D.C.: U.S. Government Printing Office, 1974); U.S. Bureau of the Census, "Geographic Mobility: March 1975 to March 1977," Current Population Reports, Series P-20, No. 32 (Washington, D.C.: U.S. Government Printing Office, 1978); National Center for Education Statistics, The Condition of Education, 1977 (Washington, D.C.: U.S. Government Printing Office, 1977).

mean percentage correct. The national mean is then subtracted from the group mean to obtain the group's difference from the national percentage correct. For example, the mean percentage of correct responses for Northeastern 17-year-olds in 1977 was 46.5. Subtracting the national mean from the Northeastern mean yields a Northeastern 17-year-old relative performance advantage of 2.3 percentage points.

Differences in group percentage (relative performance) and changes in those differences from 1969-70 to 1972-73 and from 1972-73 to 1976-77 for the full exercise sets are contained in Tables 3-1 through 3-3 for ages 9, 13 and 17, respectively.

In this report, we have chosen to emphasize changes in relative performance for several reasons. Most reporting groups changed very little in relative position over the course of the three assessments. That is, whatever the initial advantage or disadvantage of a reporting group, the average percentage of correct responses changed at about the same rate as the nation for each age population. The mean difference from the nation, since it removes the overall national trend, makes it easier to detect those reporting groups, such as extreme rural, that have undergone major shifts in position relative to the nation. Those differences were highly stable over the three assessments, as depicted in Figures 3-1 and 3-2.

Figure 3-1 shows the range of group differences from the nation at each age for sex, race, region and level of parental education. Figure 3-2 shows the same information for type of community, size of community and grade in school. For each age and reporting group, the dot is the weighted average of mean group differences from 1969-77, while a line is drawn between the most extreme mean group differences. When a consistent trend exists across the change measures, an arrowhead has been placed on the line to indicate the direction of change.

Across the three assessments:

- Males maintained their advantage over females, and the gap increased with age from about 2 percentage points at age 9 to about 6 percentage points at age 17.

- Performance of white students was consistently higher than that of black students. Differences in performance ranged from 12 to 18 points for the three age groups.

- Performance of students in the Northeast was consistently high while performance in the Southeast was consistently low, ranging from about 4 percentage points below the nation at age 9 to about 2 percentage points below at age 17. In the Central region, performance was consistently above the nation, while students in the West performed at or below the nation.

- Level of parental education was consistently related to achievement. Students reporting that neither parent graduated from high school

TABLE 3-1. Reporting-Group Mean Differences in Percentage Correct From the Nation for 1970, 1973 and 1977; Change in Mean Differences From 1970 to 1973 and From 1973 to 1977; and Standard Errors for the Total Change Exercise Sets at Age 9

	Mean Differences on 92 Exercises			Mean Differences on 71 Exercises		
	1970	1973	Change From 1970-73	1973	1977	Change From 1973-77
Sex						
Male	1.1*	1.0*	-0.1	1.3*	1.3*	0.0
Standard error	(.2)	(.2)	(.2)	(.2)	(.2)	(.2)
Female	-1.1*	-1.0*	0.1	-1.3*	-1.4*	0.0+
Standard error	(.2)	(.1)	(.2)	(.2)	(.2)	(.2)
Race						
White	3.0*	3.0*	0.0	2.7*	2.4*	-0.3
Standard error	(.2)	(.3)	(.4)	(.2)	(.3)	(.4)
Black	-14.2*	-13.6*	0.5	-12.5*	-12.8*	-0.3
Standard error	(.7)	(.6)	(.9)	(.6)	(.7)	(.9)
Region						
Northeast	2.6*	1.8*	-0.8	1.3*	2.1*	0.8
Standard error	(.5)	(.6)	(.8)	(.6)	(.6)	(.9)
Southeast	-5.8*	-4.3*	1.5	-3.8*	-4.2*	-0.4
Standard error	(.7)	(.9)	(1.1)	(.8)	(.9)	(1.1)
Central	1.7*	1.6*	-0.1	1.6*	1.1	-0.5
Standard error	(.6)	(.8)	(1.0)	(.7)	(.8)	(1.1)
West	0.4	0.3	-0.1	0.4	0.4	0.0
Standard error	(.7)	(.8)	(1.0)	(.7)	(.7)	(1.0)
Parental education						
Not graduated high school	-6.9*	-5.2*	1.8*	-5.2*	-6.4*	-1.2
Standard error	(.6)	(.5)	(.8)	(.5)	(.6)	(.8)
Graduated high school	0.5	0.7*	0.2	0.7*	1.1*	0.4
Standard error	(.4)	(3)	(.5)	(.3)	(.3)	(.4)
Post high school	5.9*	5.4*	-0.5	5.2*	4.5*	-0.7
Standard error	(.3)	(.2)	(.4)	(.3)	(.3)	(.4)
Type of community						
Extreme rural	-3.7*	-2.6*	1.0	-2.2*	0.7	2.9*
Standard error	(1.3)	(1.0)	(1.6)	(.9)	(1.1)	(1.4)
Low metro	-15.2*	-13.4*	1.8	-12.0*	-11.2*	0.8
Standard error	(1.1)	(.8)	(1.4)	(.7)	(1.3)	(1.5)
High metro	8.1*	6.6*	-1.5	5.7*	7.3*	1.6
Standard error	(.7)	(.8)	(1.1)	(.9)	(.8)	(1.2)
Size of community						
Big city	-3.6*	-3.5*	0.1	-3.6*	-4.6*	-1.0
Standard error	(.7)	(.8)	(1.0)	(.7)	(1.0)	(1.2)
Fringes around big cities	4.1*	2.8*	-1.3	2.5*	4.2*	1.7
Standard error	(.7)	(.7)	(1.0)	(.7)	(.6)	(.9)
Medium city	1.2*	1.3	0.2	2.5*	-0.7	-3.2*
Standard error	(.6)	(1.1)	(1.3)	(.9)	(1.4)	(1.6)
Smaller places	-0.1	0.0	0.1	-0.1	0.1	0.2
Standard error	(.4)	(.5)	(.6)	(.5)	(.5)	(.7)
Grade in school						
3	-9.0*	-8.4*	.6	-7.6*	-6.9*	.7
Standard error	(.5)	(.4)	(.6)	(.4)	(.4)	(.6)
4	3.4*	2.7*	- .7*	2.4*	2.3*	- .2
Standard error	(.2)	(.2)	(.2)	(.2)	(.2)	(.2)

*Denotes differences or changes in differences greater than or equal to two standard errors.
†All computations were performed prior to rounding to one decimal place.

TABLE 3-2. Reporting-Group Mean Differences in Percentage Correct From the Nation for 1969, 1972 and 1976; Change in Mean Differences From 1969 to 1972 and From 1972 to 1976; and Standard Errors for the Total Change Exercise Sets at Age 13

	Mean Differences on 67 Exercises			Mean Differences on 75 Exercises		
	1969	1972	Change From 1969-72	1972	1976	Change From 1972-76
Sex						
Male	2.0*	2.1*	0.1	1.8*	2.3*	0.5*
Standard error	(.2)	(.2)	(.3)	(.2)	(.2)	(.2)
Female	-1.9*	-2.1*	-0.2	-1.8*	-2.2*	-0.4*
Standard error	(.2)	(.2)	(.3)	(.2)	(.2)	(.2)
Race						
White	3.1*	3.3*	0.3†	2.7*	2.6*	-0.1
Standard error	(.3)	(.3)	(.4)	(.3)	(.3)	(.4)
Black	-15.2*	-16.6*	-1.4	-13.4*	-11.8*	1.6
Standard error	(.5)	(.6)	(.8)	(.5)	(1.0)	(1.1)
Region						
Northeast	2.1*	2.0*	-0.1	1.5*	2.1*	0.6
Standard error	(.6)	(.8)	(1.0)	(.6)	(.7)	(1.0)
Southeast	-4.4*	-3.2*	1.2	-2.7*	-2.7*	0.0
Standard error	(.9)	(.8)	(1.2)	(.7)	(.6)	(.9)
Central	2.2*	1.8*	-0.4	1.5*	1.6*	0.1
Standard error	(.6)	(.8)	(1.0)	(.7)	(.7)	(1.0)
West	-0.2	-0.8	-0.5	-0.4	-1.4*	-1.1
Standard error	(.6)	(.8)	(1.0)	(.7)	(.7)	(1.0)
Parental education						
Not graduated high school	-7.4*	-7.1*	0.4	-5.9*	-6.2*	-0.3
Standard error	(.5)	(.5)	(.7)	(.5)	(.6)	(.7)
Graduated high school	-1.3*	-0.3	1.1*	-0.2	-0.6*	-0.5
Standard error	(.3)	(.3)	(.4)	(.3)	(.3)	(.4)
Post high school	6.0*	6.3*	0.3	5.2*	4.9*	-0.2
Standard error	(.3)	(.3)	(.4)	(.2)	(.2)	(.3)
Type of community						
Extreme rural	-4.3*	-2.0	2.3	-1.9	-0.4	1.6
Standard error	(1.2)	(1.2)	(1.7)	(1.1)	(.9)	(1.4)
Low metro	-11.9*	-13.1*	-1.2	-10.7*	-11.6*	-0.8
Standard error	(1.1)	(1.3)	(1.7)	(1.2)	(1.4)	(1.8)
High metro	6.4*	6.8*	0.4	5.4*	5.6*	0.1
Standard error	(.8)	(.6)	(1.0)	(.6)	(.6)	(.8)
Size of community						
Big city	-3.5*	-3.8*	-0.3	-3.1*	-3.2*	-0.1
Standard error	(.8)	(.9)	(1.2)	(.8)	(1.0)	(1.3)
Fringes around big cities	3.1*	2.0*	-1.1	1.5*	2.5*	1.0
Standard error	(.6)	(.6)	(.9)	(.6)	(1.0)	(1.2)
Medium city	0.8	0.3	-0.5	0.1	-0.1	-0.2
Standard error	(1.0)	(1.2)	(1.6)	(1.1)	(1.0)	(1.5)
Smaller places	0.0	0.1	0.7	0.6	0.2	-0.4
Standard error	(.5)	(.4)	(.7)	(.4)	(.4)	(.6)
Grade in school						
7	-7.1*	-7.2*	- .1	-5.6*	-6.0*	- .4
Standard error	(.5)	(.5)	(.7)	(.4)	(.3)	(.5)
8	3.3*	3.2*	- .1	2.5*	2.4*	- .1
Standard error	(.2)	(.2)	(.3)	(.2)	(.2)	(.2)

*Denotes differences or changes in differences greater than or equal to two standard errors.
†All computations were performed prior to rounding to one decimal place.

TABLE 3-3. Reporting-Group Mean Differences in Percentage Correct From the Nation for 1969, 1973 and 1977; Changes in Mean Differences From 1969 to 1973 and From 1973 to 1977; and Standard Errors for the Total Change Exercise Sets at Age 17

	Mean Differences on 64 Exercises			Mean Differences on 70 Exercises		
	1969	1973	Change From 1969-73	1973	1977	Change From 1973-77
Sex						
Male	3.0*	2.8*	-0.2	3.5*	3.2*	-0.3
Standard error	(.2)	(.2)	(.3)	(.4)	(.4)	(.6)
Female	-2.9*	-2.7*	0.2	-3.3*	-3.2*	0.2†
Standard error	(.2)	(.2)	(.3)	(.2)	(.2)	(.3)
Race						
White	1.6*	1.9*	0.3	2.2*	2.2*	0.0
Standard error	(.2)	(.2)	(.3)	(.2)	(.2)	(.3)
Black	-11.1*	-10.4*	0.7	-12.6*	-13.5*	-0.9
Standard error	(.7)	(.4)	(.8)	(.5)	(.5)	(.7)
Region						
Northeast	1.9*	1.6*	-0.2	1.0	2.3*	1.3
Standard error	(.6)	(.5)	(.8)	(.6)	(.8)	(1.0)
Southeast	-3.2*	-1.6*	1.6	-2.1*	-2.2*	-0.1
Standard error	(.6)	(.6)	(.9)	(.7)	(.7)	(1.0)
Central	0.3	0.6	0.3	1.0	1.2*	0.2
Standard error	(.5)	(.6)	(.8)	(.6)	(.6)	(.9)
West	0.2	-1.1*	-1.3	-0.4	-1.0	-0.6
Standard error	(.5)	(.6)	(.8)	(.6)	(.7)	(1.0)
Parental education						
Not graduated high school	-5.7*	-6.3*	-0.6	-6.6*	-6.9*	-0.3
Standard error	(.4)	(.4)	(.6)	(.4)	(.4)	(.6)
Graduated high school	-1.1*	-1.4*	-0.3	-1.7*	-2.0*	-0.3
Standard error	(.3)	(.3)	(.4)	(.3)	(.2)	(.4)
Post high school	4.2*	4.2*	0.0	4.7*	4.6*	-0.2
Standard error	(.2)	(.2)	(.3)	(.2)	(.2)	(.3)
Type of community						
Extreme rural	-2.9*	-1.4*	1.6	-0.8	-0.3	0.5
Standard error	(1.0)	(.8)	(1.3)	(.8)	(.9)	(1.2)
Low metro	-5.1*	-7.3*	-2.3	-8.1*	-10.1*	-2.1
Standard error	(1.1)	(1.1)	(1.5)	(1.0)	(1.3)	(1.6)
High metro	5.9*	4.4*	-1.5	4.7*	4.4*	-0.3
Standard error	(.7)	(.8)	(1.0)	(.8)	(1.3)	(1.5)
Size of community						
Big city	-1.8*	-3.3*	-1.5	-3.6*	-4.4*	-0.8
Standard error	(.8)	(.8)	(1.2)	(1.0)	(1.0)	(1.4)
Fringes around big cities	2.1*	1.3	-0.8	1.1	2.5*	1.4
Standard error	(.7)	(.8)	(1.0)	(.8)	(.7)	(1.1)
Medium city	0.7	-0.1	-0.8	-0.1	0.2	0.3
Standard error	(.8)	(.8)	(1.2)	(.8)	(1.4)	(1.6)
Smaller places	-0.5	0.5	1.0	0.8*	0.4	-0.3
Standard error	(.4)	(.4)	(.6)	(.4)	(.5)	(.6)
Grade in school						
10	-7.5*	-7.4*	.1	-7.8*	-7.8*	0.0
Standard error	(.5)	(.6)	(.5)	(.5)	(.4)	(.6)
11	1.1*	1.1*	0.0	1.2*	1.2*	0.0
Standard error	(.1)	(.1)	(.2)	(.2)	(.1)	(.2)
12	3.1*	2.6*	- .4	2.4*	2.9*	0.6
Standard error	(.4)	(.4)	(.6)	(.4)	(.4)	(.6)

*Denotes differences or changes in differences greater than or equal to two standard errors.
† All computations were performed prior to rounding to one decimal place.

FIGURE 3-1. Sex, Race, Region and Parental-Education Reporting-Group Mean Differences in Percentage Correct From the Nation in Three Assessments, Ages 9, 13 and 17

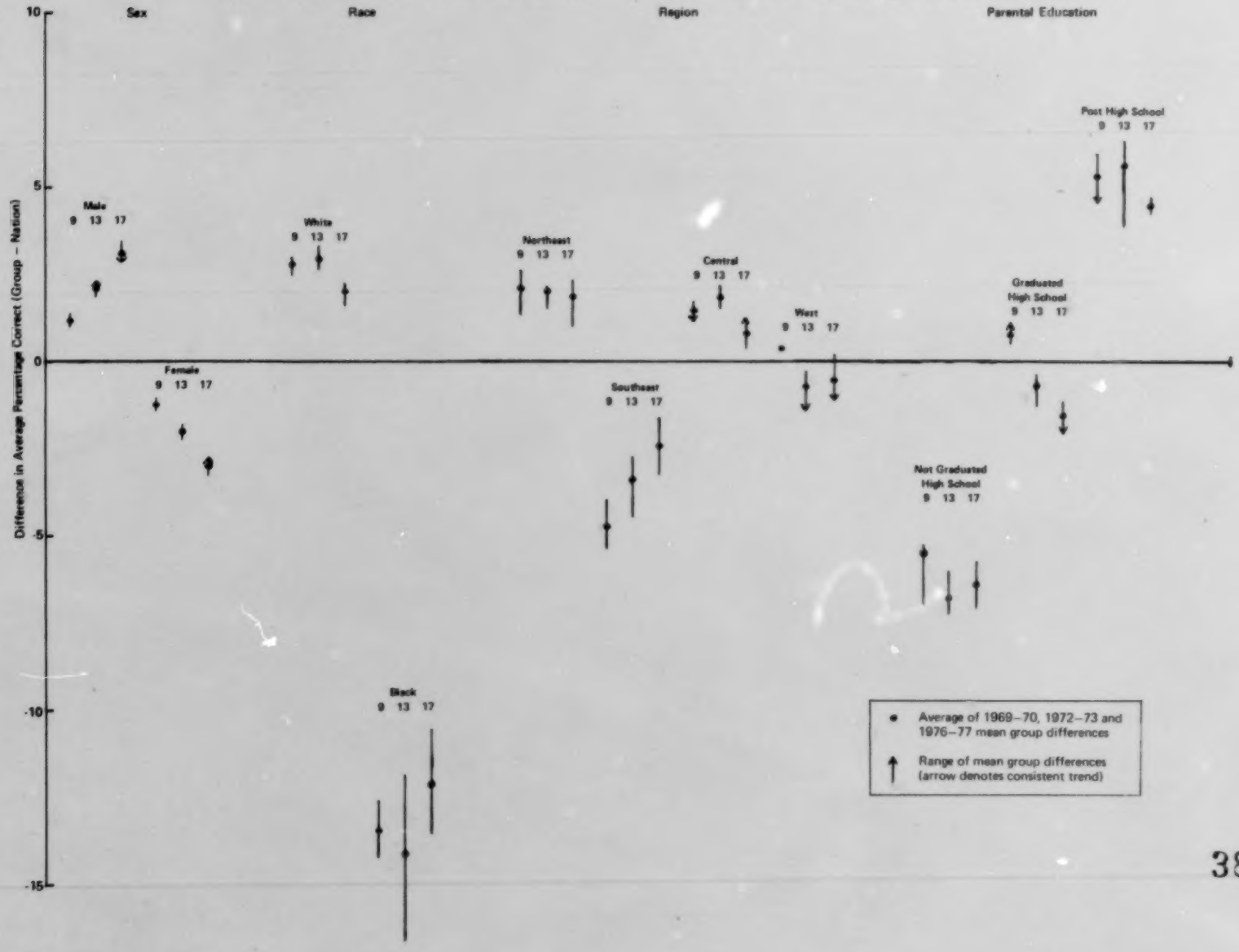
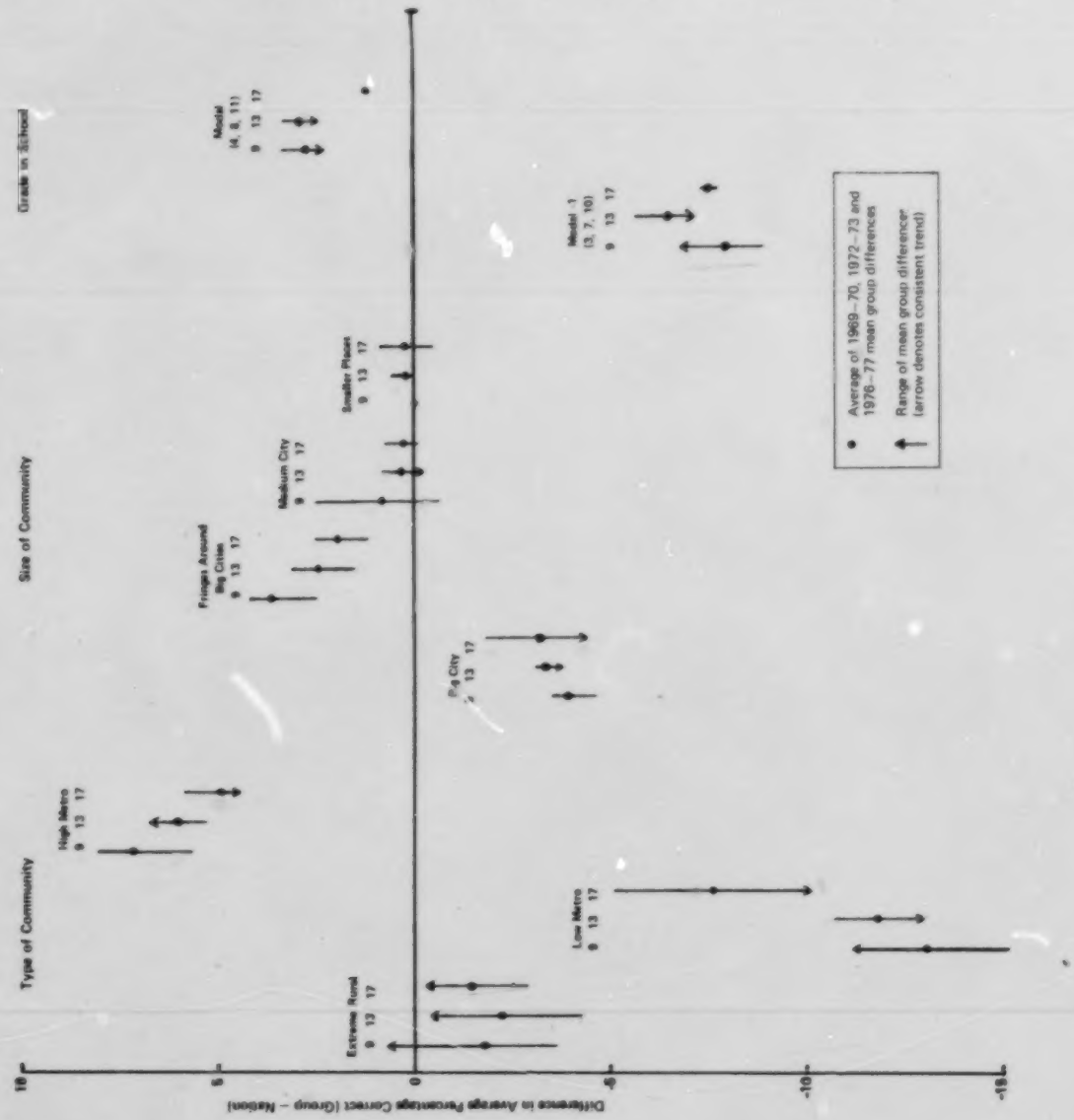


FIGURE 3-2. Type-of-Community, Size-of-Community and Grade-in-School Reporting-Group Mean Differences in Percentage Correct From the Nation in Three Assessments, Ages 9, 13 and 17



consistently achieved 5 to 7 percentage points below the nation, while those reporting at least one parent with post high school education scored 4 to 6 points above the nation.

- Students in disadvantaged-urban (low-metro) communities performed 5 to 15 percentage points below the nation, and 17-year-olds' performance was generally closer to the nation than the other ages. Students from advantaged-urban (high-metro) communities performed about 4 to 8 points above the nation; 17-year-olds, again, were closest to the nation. Students in extreme-rural areas moved from well below the nation in 1969-70 to the national level in 1976-77.

- Students in big cities consistently performed below the nation, while students in fringes around big cities performed above the nation. Performance of students in medium cities and smaller places tended to be at or near the national level, although the performance of 9-year-olds in medium cities was somewhat erratic.

- Students one grade below the modal grades for their age (grades 3, 7 and 10) consistently performed 5 to 9 percentage points below the nation, while those in the modal grades (grades 4, 8 and 11) performed 2 to 3 points above at ages 9 and 13. Seventeen-year-olds in the 11th grade were about 1 point and 12th graders 2 to 3 points above the nation.

The highly consistent performance patterns shown in Tables 3-1 to 3-3 and Figures 3-1 and 3-2 and the general lack of change in performance relative to the nation occurred through all of the science summaries. For that reason, group results for type of science, 1972-73 objectives and the exercises used in all three assessments have not been reproduced in this report.

CHAPTER 4

THE ADULT SCIENCE ASSESSMENTS

The science assessments of young adults, ages 26-35, were similar in many respects to those of 9-, 13- and 17-year-olds. Conceptually, National Assessment extended its coverage of the American population at three stages of education (late primary, middle school and high school) to an age group where most members had completed their formal schooling. The same objectives were used, and the exercises, while originally written for 13- or 17-year-olds, were also appropriate for young adults. Nationally representative probability samples of all age groups were assessed. A school sample was used for the three school-age populations, while a household sample was used for young adults. Details of sampling and data collection for the two types of surveys are sufficiently different to merit a separate discussion of the young adult assessments.

The 1969 Assessment of Young Adults

In the summer of 1969, young adults born between July 1933 and June 1943 were assessed using the same primary sampling units (PSUs) used for the in-school 17-year-old assessment.[1] It was the first large-scale attempt to collect achievement data in a household survey, and respondents were not paid to participate in the assessment. Seventy-seven percent of the sample households were successfully screened to see if any age-eligible persons lived there, while only 57% of the age eligibles who were located agreed to participate, yielding a 44% response rate. Achievement data were reported for young adults in science, citizenship and writing. However, response rates in the 1972-73 and 1977 young adult assessments were so much higher that changes in achievement from the 1969 adult assessment have not been reported. The remainder of this chapter is devoted to the second and third adult science assessments.

The 1972-73 and 1977 Assessments of Young Adults

The second and third adult science assessments were similar in large part. Their common features are described first, followed by brief summaries of

---

[1]The 1969-70 assessment is described briefly in Appendix A and more fully in 1969-1970 Science: National Results and Illustrations of Group Comparisons, Report 1, 1969-70 Assessment (Denver, Colo.: Education Commission of the States, National Assessment of Educational Progress, 1970).

unique features and summaries of changes in young adults' achievement between 1972-73 and 1977.

Both assessments were conducted by experienced household-survey staffs. Great emphasis was placed on training, supervision and verification of field work. Age-eligible adults were paid $5 per package for up to four packages of assessment exercises. As a result, nearly all sample households were successfully screened to locate age-eligible adults, and 79 to 84% of eligibles participated in the two assessments.

Sample Design

Deeply-stratified, multistage probability sample designs were used in both assessments. Stratification variables included geographic region as well as measures of community size, and urban-rural and socioeconomic-status variables.[2] Primary sampling units were made up of counties or groups of contiguous counties with 1970 census populations of at least 20,000 persons. In states that have no county definition (such as Alaska and some New England states), PSUs were defined from comparable census or political units. Table 4-1 contains the number of PSUs in each assessment.

Within each sample PSU, smaller secondary or tertiary units, or segments, were defined and sampled. Segments are small, well-identified land areas containing an average of 15 housing units in 1972-73 and 26 housing units in 1977.[3] They can range in size from one side of one block in a large city to most of a county in a rural area (number of segments used is shown in Table 4-1). Within each sample segment, all housing units were listed. A sample (sometimes 100%) of housing units was then screened for eligible adults.[4]

---

[2]The designs for 1972-73 and 1977 differed somewhat in the ways these variables were defined and the sampling stage at which stratification was introduced. In 1977, recent screening data allowed stratification of the household by race and age eligibility. Detailed documentation of the design used from 1970-71 to 1972-73 is contained in R. Moore et al., The National Assessment Approach to Sampling (Denver, Colo.: Education Commission of the States, National Assessment of Educational Progress, 1974). Similar documentation of the 1977 sample design is contained in C. Benrud et al., Final Report on National Assessment of Educational Progress Sampling and Weighting Activities for Assessment Year 08 (Raleigh, N.C.: Research Triangle Institute, February 1978).

[3]Census-defined institutions and group quarters were excluded from the definition of housing units.

[4]Adults outside the defined birth-date range, those with language barriers and those too functionally handicapped to respond to assessment materials were excluded. Self-identified nonreaders were also excluded. About 4% of age eligibles were excluded for those reasons, as shown in Table 4-2.

TABLE 4-1. Characteristics of the 1972-73 and 1977 Young Adult Assessments

	1972-73 Assessment	1977 Assessment
Subject areas	Science and mathematics	Science, energy, health and reading
Number of packages	8, both areas in each package	4, one area per package
Type of exercises	6 packages: multiple-choice and open-ended exercises 2 packages: interview/performance	Multiple-choice
Audiotapes	6 packages: self-paced 2 packages: none	Science change exercises: self-paced All other exercises: none
Incentive	$5.00 per package	$5.00 per package
Data collection period	October 1972 to May 1973	May 1977 to July 1977
Birth-date range	January 1, 1937, to December 31, 1946	January 1, 1941, to December 31, 1950
Number of primary sampling units	106	58
Number of segments	1,059	429

Field interviewers made multiple visits, if necessary, to housing units to obtain screening information from occupants or neighbors. Occupants who refused to supply screening data were called or visited by supervisory staff to solicit cooperation.

Data Collection

When eligible adults were located, they were asked to fill out a background questionnaire and complete up to four packages of assessment exercises administered in random order. If they agreed to participate, they were paid $5 for each package completed. Each package was designed to take about 45 minutes. The average number of packages completed per respondent was 3.83 in the second assessment and 3.74 in the third. Response rates were 84% and 79% in the two assessments, as shown in Table 4-2. Field work was continuously monitored by a combination of mail, telephone and personal follow ups with respondents to verify that they had been assessed and that procedures had been properly followed.

Scoring

Sets of background questionnaires and packages were audited for completeness and consistency and scored by the Measurement Research Center in Iowa City, Iowa. Sampling weights were computed and adjusted for nonresponse.

Differences Between the 1972-73 and 1977 Assessments

Sample Design

The 1972-73 sample was designed specifically for National Assessment, with large enough samples (about 2,100 per package) to allow reporting by standard assessment reporting categories (region, sex, race, parental education, and size and type of community).

The 1977 assessment of young adults used a half-sample of Research Triangle Institute's National General Purpose Sample. Because the sample was not specifically designed for National Assessment, stratification was not strictly optimal for NAEP's purposes. For example, census regional definitions were used to stratify the sample rather than the Office of Business Economics' regional definitions used by National Assessment. Also, socioeconomic indexes were used to stratify the sample, but low socioeconomic areas were not oversampled, as was done in 1972-73. The sample included 58 PSUs and about 1,300 respondents per package, which is not large enough for all National Assessment reporting categories. It was possible to oversample blacks and report results by race, but community-size categories were collapsed from four to two and neither type-of-community nor regional results are reported.

TABLE 4-2. Screening and Response-Rate Data for the 1972-73 and 1977 Assessments of Young Adults

	1972-73		1977	
	Number	Percentage	Number	Percentage
Number occupied housing units	14,395	--	2,123†	--
Number and percent screened	14,346	99.7	2,084	98.2
Listed age eligibles	5,225	--	1,831	--
Less: Language barriers	174	3.3	45	2.5
Handicapped	50	1.0	15	.8
Nonreader	††	--	24	1.3
NAEP eligibles	5,001	100.0	1,747	100.0
Respondents	4,211	84.2	1,379	78.9
Refused	733	14.7	316	18.1
Not home	57	1.1	52	3.0
Total number packages	16,112	--	5,158	--
Average packages/respondent	3.83	--	3.74	--
Average respondents/package	2,103	--	1,290	--
Number respondents for science change summaries	12,079	--	1,289	--

†*Housing units were subsampled prior to screening in 1977.*
††*Nonreader data were not separately identified in 1972-73.*

Learning-Area Mix

The 1972-73 assessment comprised eight packages, with science and mathematics exercises in each package. Many of the exercises were open-ended, and two packages contained interview and performance items. The 1977 assessment contained four packages, one each of science, reading, energy and health exercises.[5]

Taping

In 1972-73, self-paced audiotapes were used in administering exercises. The field interviewer turned on the tape recorder while the respondent read an exercise; the recorder was then turned off until the respondent was ready to go on to the next exercise. In the 1977 assessment, the same procedure was used for science exercises repeated from 1972-73. Those exercises occurred in the first half of the science package. Audiotapes were not used for any other exercises. In the 1972-73 assessment, the tapes accompanied all exercises except the interview and performance items, while in 1977, they were used for only a small portion of the total testing period.

Time of Testing

The 1972-73 assessment was conducted between October and May, while the 1977 assessment was conducted between May and July. Almost all young adults between the ages of 26 and 35 have completed their formal education and there is little reason to believe that time of testing has so much effect on their performance as it has on that of school-age populations.

It is also possible that there are some differences in the characteristics of nonrespondents between fall and spring, and between late-spring and early-summer survey periods. The shorter survey period in 1977 did make it more difficult to locate and assess eligibles who were away from home for extended periods.

Released Exercises

As explained in Appendix C, National Assessment is very cautious about using released exercises to measure changes in the achievement of school-age populations. There are many ways students can be exposed to exercises that we have released for public use. We are, however, less concerned about reusing released exercises with young adults. For adults, we have not identified any

---

[5]A short questionnaire was also administered for the Food and Drug Administration after all other packages had been completed. Ninety percent of the respondents agreed to complete this questionnaire.

plausible analogs to the ways students can be systematically exposed to released exercises. Adults are not likely to be exposed to specific exercises unless they are teachers or researchers directly involved in the subject area -- a very small proportion of the population.

Consequently, a number of previously released exercises were included in the 1977 adult assessment to measure changes in science achievement. Four of the 20 change exercises had been released after either the 1969-70 or 1972-73 assessments.

National Results

Mean percentages of correct responses for each assessment and changes in percentages of correct responses for young adults between 1972-73 and 1977 are given in Table 4-3.[6] The percentage of correct responses decreased 4 percentage points between the two assessments on the 20 exercises available to measure changes in achievement. Fifteen of the exercises were also administered to 17-year-olds enrolled in school during the two science assessments. The percentages of correct responses were similar for 17-year-olds and adults in the two assessments, and both age groups' percentage correct decreased between 1972-73 and 1977.

Mean percentages of correct response and changes for both released and unreleased exercises are also shown in Table 4-3. Decreases for both sets were about 4 percentage points, the same as for the entire set of exercises.

Group Results for Young Adults

Differences in mean percentages of correct responses between reporting groups and the nation in 1972-73 and 1977 and changes in those differences are displayed in Table 4-4. Table 4-4 also contains estimated population proportions for each reporting group and estimated standard errors for all mean differences and changes in mean differences. Adult reporting-group definitions are identical to those given in Chapter 3, with the following exceptions:

- Community-size categories were collapsed. Big cities and fringes around big cities were combined, as were medium cities and smaller places.
- Young adults' own education is reported, using the same category definitions as parental education for students in school.
- The age range 26-35 was divided into ages 26-30 and 31-35.

---

[6]Unlike the results for 9-, 13- and 17-year-olds, young adults' weights were not smoothed prior to estimating proportions and percentages of correct responses. Data were available for only the assessment years 1970-71 through 1973-74 and 1977 -- too few points for effective weight smoothing.

TABLE 4-3. Mean Percentages Correct, Changes in Mean Percentages Correct and Standard Errors From 1972-73 to 1977 for Young Adults and for 17-Year-Olds on Selected Exercises

	Number of Exercises	Mean Percentage Correct			Standard Errors of Mean Percentage Correct		
		1972-73	1977	Change	1972-73	1977	Change
All exercises	20	44.6	40.7	-3.9*	.6	1.0	1.2
Released exercises	4	54.8	50.8	-4.1	†	†	†
Unreleased exercises	16	42.0	38.2	-3.9	†	†	†
Exercises administered to both 17-year-olds and adults							
Adults	15	40.6	36.8	-3.8*	.6	1.0	1.2
17-year-olds	15	41.9	36.4	-5.6*	.5	.6	.7
Differences between adults and 17-year-olds	15	-1.4	.4	1.8	.8	1.2	1.4

Denotes changes in mean percentages greater than or equal to two standard errors.
†*Standard errors not available.*

TABLE 4-4. Estimated Population Proportions, Mean Differences in Percentage Correct From the Nation for Young Adult Reporting Groups Between 1972-73 and 1977, Changes in Mean Differences and Standard Errors

	Estimated Population Proportions		Mean Differences			Standard Errors of Mean Differences		
	1972-73	1977	1972-73	1977	Change	1972-73	1977	Change
Sex								
Male	.48	.47	7.7*	8.4*	.7	.4	.7	.8
Female	.52	.53	-7.1*	-7.3*	- .2	.4	.7	.8
Race								
White	.84	.81	3.2*	4.0*	.8	.3	.8	.9
Black	.10	.13	-18.7*	-19.2*	- .5	1.1	1.2	1.6
Other	.06	.06	--	--	--	--	--	--
Own education								
Not graduated high school	.21	.10	-17.4*	-20.1*	-2.7*	.8	1.1	1.4
Graduated high school	.33	.30	-6.0*	-6.0*	.0	.7	1.0	1.2
Post high school	.46	.52	12.3*	10.6*	-1.7	.5	.8	1.0
Parental education								
Not graduated high school	.41	.37	-7.7*	-8.2*	- .5	.5	1.1	1.2
Graduated high school	.30	.29	2.7*	1.2	-1.5	.7	1.2	1.3
Post high school	.25	.29	11.4*	11.8*	.5	.8	1.4	1.6
Community size								
Big city and fringes	.42	.39	- .2	.5	.8	.7	1.4	1.6
Medium city and smaller places	.58	.61	.2	- .3	- .5	.5	.9	1.0
Age								
26-30	.57	.55	.6	1.5*	.9	.3	.7	.8
31-35	.43	.45	- .8	-1.8*	-1.0	.5	.8	1.0

**Denotes differences or changes in differences greater than or equal to two standard errors.*

Results for young adults paralleled results for the school-age populations reported in Chapter 3. There were large differences in performance between males and females, whites and blacks, and among levels of education. The relationship between education level and achievement was generally stronger for own education than parental education. By age group, 26- to 30-year-olds generally performed better than 31- to 35-year-olds, although only the 1977 differences were significant.

Changes in performance for most reporting groups were similar to the decline in performance for all adults between 1972-73 and 1977. Adults who had not graduated from high school in 1973-74 performed 17 percentage points below the nation and in 1977, 20 points below the nation.

APPENDIX A

TECHNICAL PROCEDURES: SAMPLING AND ESTIMATION OF STANDARD ERRORS

This appendix describes technical procedures used in the sample design and sample selection for the 1969-70, 1972-73 and 1976-77 assessments of science exercises. It also provides details about procedures used for estimating standard errors of the statistics used in this report.

The sample design, sample selection and data collection for all three assessments of science were done by the staff of the Research Triangle Institute, Raleigh, North Carolina. In the first two science assessments, data collection in the Central and Western regions was subcontracted to Westinghouse Learning Corporation/Measurement Research Center, Iowa City, Iowa.

Sampling

The target populations for each of the assessments consisted of 9-, 13- and 17-year-olds enrolled in either public or private schools at the time of the assessment who were not functionally handicapped to the extent that they could not participate in an assessment. Specific groups excluded were: non-English speaking persons, those identified as nonreaders during the assessment, persons physically or mentally unable to respond, and persons in institutions or attending schools established for the physically handicapped or mentally retarded.

Packages of exercises (item booklets) were administered to groups of students in sessions that could not exceed 50 minutes. Specially trained field personnel administered the packages. These procedures ensured a relatively high degree of consistency of administration and quality control of field procedures with minimum impact or burden on the cooperating schools.

The definitions of the target populations, exercise formats and administration procedures were identical in each assessment. The sample design used to obtain representative samples of the target populations was modified somewhat between assessments. An overview of the general sample design approach is given below, followed by a description of differences between the 1969-70, 1972-73 and 1976-77 procedures.

Overview of the National Assessment Sample Design

National Assessment uses a deeply stratified, three-stage national probability sample design with oversampling of low-income areas. In the first

stage, the United States is divided into geographical units of counties or groups of contiguous counties meeting a minimum size requirement. These units, called primary sampling units (PSUs), are stratified by region and size of community. From the list of PSUs, a sample of PSUs is drawn without replacement with probability proportional to population size measures, representing all regions and sizes of communities. Oversampling of low-income and extreme-rural areas is performed at this stage by adjusting the estimated population size measures of these areas to increase sampling rates.

In the second stage, all public and private schools within each PSU selected in the first stage are listed. Schools within each PSU are selected without replacement with probabilities proportional to the number of age eligibles in the school.

The third stage of sampling occurs during the data collection period. A list of all age-eligible students within each selected school is made. A simple random selection of eligible students, without replacement, is obtained, and item booklets are administered to selected students. Specially trained field personnel select the sample and administer the booklets. In each assessment, 13-year-olds are assessed in the months of October, November and December; 9-year-olds in January and February; and in-school 17-year-olds in March and April.

When funding levels permit, the sample of in-school 17-year-olds is supplemented with a sample of out-of-school 17-year-olds. Between 1969-70 and 1972-73, out-of-school 17-year-olds were assessed as part of the household sample of young adults. The out-of-school 17-year-old population is relatively small and expensive to locate through a household sample. Starting in 1970-71, the household sample was augmented by a supplementary sample selected from lists of dropouts and early graduates obtained from the schools at the time of the regular assessment. From 1973-74 on, only the supplementary sample has been used to assess out-of-school 17-year-olds. The household sample was dropped because it afforded only slightly better population coverage while costing much more than the supplementary one.

In 1976-77, funding limitations precluded any assessment of out-of-school 17-year-olds. In order to make the 17-year-old populations comparable for all three science assessments, results are given for 17-year-olds enrolled in public or private schools during each assessment. Results for out-of-school 17-year-olds are not included in this report.

Each respondent in the sample does not have the same probability of selection because some subpopulations are oversampled, and adjustments are made to compensate for some schools' refusal to participate and for student nonresponse. The selection probability of each individual is computed, and its reciprocal is used to weight each response in any statistical calculation to compensate for unequal rates of sampling and to ensure proper representation in the population structure.

The number of PSUs, schools within PSUs and students within schools are determined by optimum sampling principles. That is, a sample design is selected

that will minimize costs while achieving a desired level of precision.

Table A-1 displays the number of PSUs and schools within PSUs selected in 1969-70, 1972-73 and 1976-77 by age.

TABLE A-1. Number of PSUs and Schools Within PSUs Selected in 1969-70, 1972-73 and 1976-77

	1969-70 Assessment		1972-73 Assessment		1976-77 Assessment	
	No. of PSUs	No. of Schools	No. of PSUs	No. of Schools	No. of PSUs	No. of Schools
Age 9	204	935	116	971	75	451
Age 13	205	749	116	979	75	472
Age 17	193	670	116	798	75	428

Differences in Sample Design: 1969-70, 1972-73 and 1976-77

The 1976-77 sample was drawn according to the following procedures. Two types of PSUs were identified: (1) large-size population areas defined by the U.S. Bureau of the Census as Standard Metropolitan Statistical Areas (SMSAs) and (2) other contiguous non-SMSA counties grouped together to meet certain minimum-size requirements. The first stratification of PSUs was by geographic region, as defined by the Office of Business Economics, U.S. Department of Commerce (see Chapter 3).

Within regions, PSUs were classified into five size-of-community (SOC) categories:

SOC 1 — PSUs corresponding to the 13 largest SMSAs after adjusting the population size to compensate for oversampling low-income metropolitan areas. These PSUs have selection probabilities so large that under our allocation procedures they are certain to be included in our sample each year. These PSUs are designated as self-representing.

SOC 2 — PSUs corresponding to the remaining 57 SMSAs with over 500,000 population.

SOC 3 — PSUs corresponding to the remaining 162 SMSAs.

SOCs 4, 5 — PSUs made up of non-SMSA counties. SOCs 4 and 5 are determined so that half of the remaining population (after adjustment for oversampling of rural areas) falls into each category. SOC 4 contains PSUs in which less than 60% of the residents are classified as rural.

Since the self-representing PSUs are included in the sample every year, they actually represent an additional level of stratification, making an

effective total of 17 (13 + 4) size-of-community strata. Each self-representing SMSA was divided further into geographical substrata or nonoverlapping replicates that constituted multiples of convenient work units for item administration. These multiple work units were included with the rest of the non-self-representing PSUs to form the pool from which first-stage sampling units were selected. To ensure adequate representation, National Assessment doubled the sampling rate of low-income and rural areas.

In 1975-76, first-stage units were selected simultaneously for four consecutive assessment years (1975-76 through 1978-79), as were schools in the self-representing PSUs. The present sample design requires that every four years we will assess at least once in every state and not more than once in any school. There are 1,101 primary sampling units in the primary sampling frame for the four-year period, from which about 75 first-stage sampling units are selected each year.

Within the primary strata, public and private schools were listed and further stratified by the estimated number of youngsters eligible at each age. Small schools were clustered until they were large enough to respond to the same number of packages as the larger schools in a stratum. Schools or school clusters were selected without replacement with probability proportional to the number of age eligibles in the school or cluster of schools. Once schools were identified, districts were contacted to check for changes in grade range and for the existence of new schools. This information was used to revise probabilities of schools' selection.

In the third stage, students were selected with equal probability and without replacement within each sampled school. The number of students selected was proportional to the number of age eligibles, with oversampling in low-income and rural areas.

During data collection, allowing for variable group sizes for each package administration within schools enabled National Assessment to obtain desired sample sizes in schools having characteristically low response rates. This feature also permitted last-minute modifications and adjustments to selection probabilities necessitated by enrollment changes.

The sampling procedures used in 1969-70 and 1972-73 differed somewhat from those used in the 1976-77 assessment.[1] First, size measures for SMSAs, counties

---

[1]For details on the 1969-70, 1972-73 and 1976-77 sample design and data collection procedures, respectively, see C. Benrud et al., Final Report on National Assessment of Educational Progress Sampling and Weighting Activities for Assessment Year 08 (Research Triangle Park, N.C.: Research Triangle Institute, 1977); J. Chromy and D. Horvitz, "Structure of Sampling and Weighting," 1969-1970 Science: National Results and Illustrations of Group Comparisons, Report 1, 1969-70 Assessment (Denver, Colo.: Education Commission of the States, National Assessment of Educational Progress, 1970); R. Moore et al., The National Assess-

and urban areas in 1976-77 were based on 1970 census data, while those in 1969-70 were based on 1960 census data. Size measures in 1971-72 were based on 1960 census data and first-count data from the 1970 census.

Another difference occurred in the PSU sample design. In 1969-70, PSUs were stratified by region, size of community, a measure of socioeconomic status (SES) and geographic proximity. There was no requirement that all states be included in the sample. In 1972-73, the PSUs were stratified by region, size of community and SES. In addition, the sample was constrained to include all states. The sampling of PSUs in 1972-73 was accomplished by using a controlled selection procedure.[2] In 1976-77, PSUs were stratified by region and size of community, with the constraint that each state must appear in the sample once every four years and controlled selection of PSUs be abandoned.

The size-of-community (SOC) stratifications in 1969-70 and 1972-73 were similar to each other but different from those of 1976-77. There were only four SOC stratifications in the first assessment of science. The first SOC category in 1969-70 and 1972-73 consisted of all central cities with overall population greater than 180,000. The second SOC category consisted of the remainder of the SMSA containing the central city in SOC 1. The SOC 3 category in 1969-70 consisted of the remainder of the SMSAs and all counties not in SOCs 1 and 2, containing at least one city with a population over 15,000. SOC 3 for 1972-73 was similar, except that the minimum population of the city was 25,000. In both 1969-70 and 1972-73, the SOC 4 category consisted of all the remaining counties not in SOCs 1, 2 or 3.

In 1976-77, oversampling of low-income metropolitan areas and extreme-rural areas was accomplished at the primary stage by increasing the estimated population size measures of PSUs containing these areas and then sampling with probabilities proportional to these adjusted size measures. In 1969-70 and 1972-73, a poverty index was used to stratify PSUs into high- and low-SES stratifications. The sampling rates within these strata were then increased in order to achieve the desired oversampling.

In the 1976-77 assessment, packages of exercises were administered in schools to groups of students varying in size depending on an estimate of the rate of nonresponse for that school. The administration session sizes were planned to vary from 10 to about 35 students at each age. In 1969-70 and 1972-73, the planned session sizes were fixed at 12 students at each age.

---

ment Approach to Sampling (Denver, Colo.: Education Commission of the States, National Assessment of Educational Progress, 1974).

[2]R. Moore et al., The National Assessment Approach to Sampling.

Estimation of Standard Errors

Several measures of achievement that National Assessment uses in its reports were described in Chapter 1. The sample designs described in the previous section are complex, deeply stratified, multistage probability sample designs. A reasonably good approximation of standard error estimates of these achievement measures can be obtained by applying the jackknife procedure to first-stage sampling units within strata, using the method of successive differences and accumulating across strata.

In this section the measures of achievement are first defined in algebraic form, followed by a description of the jackknife method used by National Assessment to estimate their standard errors.

Measures of Achievement

Based on the sample design, a weight is assigned to every individual who responds to an exercise administered in an assessment. The weight is the reciprocal of the probability of selecting a particular individual to take a particular exercise. Since the probabilities of selection are based on an estimated number of people in the target age population, the weight for an individual estimates the number of similar people that that individual represents in the age population.

A sum of the weights for all individuals at an age level responding to an exercise is an estimate of the total number of people in that age population. A sum of weights for all individuals at an age responding correctly to an exercise is an estimate of the number of people who would be able to respond correctly in the age population, if the entire population were assessed. These concepts also apply to any reporting group (e.g., defined by region, sex, race, etc.) and category of response (e.g., correct, incorrect and "I don't know).

Let W^e_{ihk} = sum of weights for respondents to exercise e who are in reporting subgroup i who are in the kth PSU of the hth sampling stratum.

C^{ej}_{ihk} = sum of weights for respondents to exercise e who are in subgroup i, who are in the kth PSU of stratum h and who selected response category j (e.g., correct response) for the exercise.

Note that $W^e_{ihk} = \sum_j C^{ej}_{ihk}$.

Then, summing k over the n_h, sample PSUs in stratum h, and summing over the H sampling strata, $W^e_{i++} = \sum_{h=1}^{H} \sum^{n_h}_{k=1} W^e_{ihk}$ estimates the number of eligibles in the

population who are in subgroup i.

Similarly, $C^{ej}_{i++} = \sum_{h=1}^{H} \sum_{k=1}^{n_h} C^{ej}_{ihk}$ estimates the number of eligibles in the population who are in subgroup i and who would select response category j for exercise e:

$$(1) \quad p^{ej}_{i} = C^{ej}_{i++} / W^{e}_{i++}.$$

In the special case where the percentage of all age eligibles who would select response category j on exercise e is estimated, the index A (for All) will be used in place of i as follows:

$$(2) \quad p^{ej}_{A} = C^{ej}_{A++} / W^{e}_{A++}.$$

In National Assessment reports, the proportion in (1) multiplied by 100 is called the group percentage, and the proportion in (2) multiplied by 100 is called the national percentage. The difference between the proportion in subgroup i who would select category j on exercise e and the proportion in the nation is denoted by:

$$(3) \quad \Delta p^{ej}_{i} = p^{ej}_{i} - p^{ej}_{A}.$$

National Assessment also reports the arithmetic mean of the percentage of correct responses over sets of exercises corresponding to the measures in (1), (2) and (3). These means are taken over the set of all exercises or a subset of exercises classified by a reporting topic or content objective. The mean percentage of correct responses taken over m exercises in some set of exercises corresponding to measures (1), (2) and (3) are, respectively:

$$(4) \quad \overline{P}_{i} = \frac{1}{m} \sum_{e} C^{e}_{i++} / W^{e}_{i++},$$

$$(5) \quad \overline{P}_{A} = \frac{1}{m} \sum_{e} C^{e}_{A++} / W^{e}_{i++} \text{ and}$$

$$(6) \quad \overline{\Delta P}_{i} = \overline{P}_{i} - \overline{P}_{A}.$$

Note that the response category subscript j has been suppressed since the means are understood to be taken over the correct response category for each exercise.

Each of these six achievement measures are computed and routinely used in reports describing achievement data for any assessment. The simple differ-

ence in these measures between two assessments of the same exercises (or sets of exercises) provides six measures of change in achievement that are routinely used in National Assessment's change reports. The next section describes how standard errors are estimated for the 12 statistics routinely used in National Assessment reports.

Computation of Standard Errors

In order to obtain an approximate measure of the sampling variability in the statistics (1) through (6), a jackknife replication procedure for estimating the sampling variance of nonlinear statistics from complex, multistage samples was tailored to National Assessment's sample design. References (4), (5) and (7) provide information about the jackknife technique, while reference (3) describes how the procedure is used in estimating standard errors for National Assessment's sample designs.

To demonstrate the computational aspects of this technique, consider estimating the variance of the statistic in (1) -- the proportion of age eligibles in subgroup i who would select response category j on exercise e.

This statistic is based on data from all the n_h PSUs in the H strata. Let $p^{ej}_{i\text{-}hk}$ be defined as a replication estimate of P^{ej}_i and constructed from all the PSUs, excluding the data from PSU k in stratum h. These replication estimates are computed as if the excluded PSU had not responded and a reasonable nonresponse adjustment is used to replace the data in PSU hk in estimating P^{ej}_i. Several choices for replacing the data in PSU hk are available. In order to obtain a convenient and computationally efficient algorithm for approximating standard errors, National Assessment replaces C^{ej}_{ihk} and W^{e}_{ihk} from the hk<u>th</u> PSU with corresponding sums from another paired PSU in the same stratum. The replicate estimate is then computed. The replicate estimates to be used in the calculations are determined by arranging all of the PSUs in each stratum into successive pairs. That is, PSU 1 is paired with PSU 2, PSU 2 with PSU 3, 3 with 4, ... $(N_h\text{-}1)$ with n_h and PSU n_h with PSU 1.

The contribution to the variance of P^{ej}_i by each pair of PSUs is the change in the value of the statistic incurred by replacing the data from each PSU in

the pair with the data from the other PSU in the pair and recomputing P_i^{ej} in the usual way. This produces two replicate estimates. Squaring the difference between these replicate estimates and then dividing by eight measures the contribution of this pair of PSUs to the total variance. The sum of these contributions over all n_h successive pairs in the stratum is the contribution by stratum h to the total variance. The square root of the sum of the H stratum contributions is the estimate of the standard error of P_i^{ej}.

Algebraically, the two replicate estimates for the pair k, k+1 (where k=1, ..., n_h and $n_h+1 = 1$) are:

$$(7)\quad P^{ej}_{i\text{-}hk(k+1)} = \frac{C^{ej}_{i++} - C^{ej}_{ihk} + C^{ej}_{ih(k+1)}}{W^{e}_{i++} - W^{e}_{ihk} + W^{e}_{ih(k+1)}}$$

and

$$(8)\quad P^{ej}_{i\text{-}h(k+1),k} = \frac{C^{ej}_{i++} - C^{ej}_{ih(k+1)} + C^{ej}_{ihk}}{W^{e}_{i++} - W^{e}_{ih(k+1)} + W^{e}_{ihk}}.$$

The contribution to the total variance from stratum h is:

$$(9)\quad \text{var}\,(P^{ej}_{ih}) = \frac{1}{8}\sum_{k}^{n_h} (P^{ej}_{i\text{-}hk(k+1)} - P^{ej}_{i\text{-}h(k+1),k})^2.$$

And, finally, an estimate of the standard error of P_i^{ej} is:

$$(10)\quad \text{SE}\,(P_i^{ej}) = (\sum_{h}^{H} \text{var}\, P^{ej}_{hk})^{\frac{1}{2}}.$$

Multiplying P_i^{ej} by 100 yields the percentage of response to category j. Multiplying SE (P_i^{ej}) by 100 yields the corresponding estimated standard error of the percentage.

In general, the jackknifed standard errors of the proportion estimates will be larger than the simple random sampling formula $(pq/n)^{\frac{1}{2}}$, where $p = P_i^{ej}$,

$q = 1-p$ and n is the number of sampled respondents in subgroup i who took the exercise. The larger size of SE (P_i^{ej}) reflects mainly the loss of precision due to cluster-sampling of schools and students.

The standard errors for the achievement measures (2) through (6) are computed through a series of steps analogous to those followed in computing SE (P_i^{ej}).

The most complicated step in computing standard errors occurs in forming the paired replicate estimates analogous to (7) and (8) for each successive pair of PSUs. Once this bookkeeping chore is done, the computations for (9) and (10) follow in a straightforward manner.

The standard errors for the differences between two assessments for any of the achievement measures (1) through (6) are computed as the square root of the sum of the squared standard errors from each of the separate assessments.

The size of the standard errors depends largely on the number of PSUs and schools included in the sample (Table A-1), but also on the number of respondents in each of the reporting groups. Table A-2 shows the average number of students responding to an exercise package for each of the reporting groups discussed in this report, for each age and for each of the three science assessments.

The size of the standard errors of the means of the achievement measures for sets of exercises is also influenced by the number of exercises in the exercise set and the number of packages over which the items in the set are spread. Tables A-3 and A-4 show the number of exercises and packages included in the mean achievement measure for each of the content categories included in this report.

TABLE A-2. Average Number of Respondents in Reporting Groups Taking a Package of Exercises, by Age and Assessment Year

	Age 9			Age 13			Age 17 In School†		
	1970	1973	1977	1969	1972	1976	1969	1973	1977
National	2,434	2,663	2,478	2,411	2,612	2,565	2,083	2,351	2,649
Region									
Northeast	618	656	585	625	651	587	577	573	581
Southeast	563	669	646	562	667	708	423	596	673
Central	574	672	736	570	649	764	507	596	850
West	678	665	510	654	645	506	577	586	545
Sex									
Male	1,231	1,328	1,245	1,166	1,294	1,268	1,013	1,126	1,313
Female	1,203	1,335	1,233	1,231	1,318	1,297	1,070	1,225	1,336
Race									
White	1,825	1,997	1,911	1,799	1,977	1,940	1,723	1,852	2,155
Black	390	466	391	416	436	473	237	358	359
Parental education									
Not graduated high school	269	271	234	361	417	336	432	455	418
Graduated high school	562	564	681	762	792	868	685	720	897
Post high school	794	787	729	1,038	994	1,013	931	1,028	1,212
Type of community									
Extreme rural	240	265	247	242	263	261	206	230	256
Low metro	243	266	252	243	264	263	212	239	266
High metro	243	267	241	239	260	261	209	234	267
Size of community									
Big city	665	619	617	651	583	633	558	439	622
Fringes around big cities	403	515	461	381	531	484	334	493	488
Medium city	326	372	274	371	365	218	334	326	274
Smaller places	1,040	1,157	1,126	1,007	1,133	1,230	857	1,094	1,265
Grade									
3, 7, 10	558	646	575	581	693	663	267	305	350
4, 8, 11	1,779	1,946	1,855	1,728	1,809	1,842	1,446	1,688	1,977
12	--	--	--	--	--	--	323	304	286

†*Seventeen-year-olds enrolled in school, excluding follow ups in the 1977 assessment.*

TABLE A-3. Number of Change Exercises in Various Content Categories and Number of Packages in Which Those Exercises Appeared in 1969-70 and 1972-73, Ages 9, 13 and 17

	Age 9			Age 13			Age 17		
	Number of Exercises	Number of Packages		Number of Exercises	Number of Packages		Number of Exercises	Number of Packages	
		1970	1973		1969	1972		1969	1973
All exercises	92	8	7	67	9	9	64	11	11
Type of science									
Biology	27	8	7	23	8	9	20	10	10
Physical science	50	8	7	36	9	9	39	10	11
Unclassified	15	8	7	8	6	7	5	5	5
1972-73 objectives									
Know	40	8	6	37	9	8	34	11	11
Understand/apply	47	8	7	28	9	9	30	10	11
Appreciate	5	5	5	2	2	2	0	0	0
Exercises used in all three assessments	30	8	6	23	8	8	23	9	10

TABLE A-4. Number of Change Exercises in Various Content Categories and Number of Packages in Which Those Exercises Appeared in 1972-73 and 1976-77, Ages 9, 13 and 17

	Age 9			Age 13			Age 17		
	Number of Exercises	Number of Packages		Number of Exercises	Number of Packages		Number of Exercises	Number of Packages	
		1973	1977		1972	1976		1973	1977
All exercises	71	8	7	75	9	10	70	11	11
Type of science									
Biology	24	7	6	23	9	10	19	10	10
Physical science	42	7	7	47	9	10	45	11	11
Unclassified	5	4	4	5	4	4	6	5	5
1972-73 objectives									
Know	32	7	7	38	9	10	31	11	11
Understand/apply	36	8	7	37	9	10	37	11	11
Appreciate	3	3	3	0	0	0	2	2	2
Exercises used in all three assessments	30	6	6	23	8	10	23	10	10

REFERENCES

Folsom, R.E., Jr. National Assessment Approach to Sampling Estimation, Sampling Error Monograph 25U-796-5. Research Triangle Park, N.C.: Research Triangle Institute, 1977.

Miller, R.G., Jr. "A Trustworthy Jackknife." Annals of Mathematical Statistics, No. 39 (1968), pp. 567-582.

Miller, R.G., Jr. "The Jackknife -- A Review." Biometrika, No. 61 (1974), pp. 1-15.

Mosteller, F., and Tukey, J.W. Data Analysis and Regression (Chapter 8). Reading, Mass.: Addison-Wesley, 1977.

APPENDIX B

ESTIMATED POPULATION PROPORTIONS OF REPORTING GROUPS BASED ON NATIONAL ASSESSMENT SAMPLES, 1969-70, 1972-73 AND 1975-76

The estimated population proportions for reporting groups shown in this appendix are based on weights derived from the sampling process used in the three assessments of 9-, 13- and in-school 17-year-olds. These proportions vary from year to year due to random sampling variability or systematic differences in sampling procedures. A better estimate of population proportions for any single year can be obtained by smoothing[1] the proportions over several assessment years. Smoothing does not make the estimated proportions identical but does reduce variability. The estimated population proportions shown in this appendix and used in estimating performance were obtained after smoothing proportions from the first eight years of assessment. The procedures used to obtain the smoothed proportions are detailed below.

The purpose of smoothing estimated population proportions is to reduce sampling fluctuations that can affect estimates of the change over time in the percentage of acceptable responses to an exercise. For example, the percentage of acceptable responses for an age group is a function of the relative proportions of high-performing and low-performing groups. If the relative proportions of these groups are very different in different assessments due to sampling variability, then a portion of the change in percentage of acceptable responses for an age group is directly attributable to yearly sampling differences in the relative proportions of high- and low-achieving groups. Smoothing estimates of population proportions reduces a large portion of the sampling variability while preserving, as far as possible, actual trends occurring in the age population.

The specific procedure used to obtain the smoothed population proportions that were used in this report is detailed below. This procedure, which was applied independently to each of the three age groups, is basically a weighting-class adjustment applied independently to each reporting category (nation, region, sex, etc.). By applying this weighting-class procedure independently

---

[1]The word "smoothing" is used here in the sense of drawing a "smooth" curve to fit a sequence of numbers. Proportions for each reporting group covering eight years were smoothed by the robust/resistant procedures described in Chapter 7, Exploratory Data Analysis by John W. Tukey (Reading, Mass.: Addison-Wesley, 1977).

to each reporting variable, it was possible to produce good estimates of the marginal proportions of people within each category of the variable, while disturbing as little as possible the relationships between other reporting variables within the adjusted variable.

The same weighting-class partitioning of the population was used for all ages and reporting variables. For each age, the entire population of eligibles was partitioned into nine cells, called smoothing cells, on the basis of membership in a variety of demographic categories determined by the race, grade, home reference items and parental-education variables. The purpose of the partitioning was to obtain subgroups of eligibles that exhibited substantial differences in performance on science exercises. In addition to differentiating on performance, the smoothing cells were required to contain adequate numbers of eligibles to ensure stability of the weight adjustments. These criteria produced the smoothing cells detailed in Table B-1.

For each age and reporting variable, the population of eligibles was partitioned into subgroups determined by the various categories of the variable and by the smoothing cells. For example, classification of the population by sex and the smoothing cells produced a partitioning consisting of 18 subgroups: males in smoothing cell 1, males in smoothing cell 2, ..., females in smoothing cell 9. Estimates of the proportions of eligibles in each of the subgroups were then obtained for each of the eight assessment years. The estimated proportion of eligibles in a particular subgroup for a given year was computed as the sum of weights of respondents in the subgroup assessed that year divided by the sum of weights of all eligibles.

This produced, for each subgroup, estimated proportions for each of the eight assessment years. Each such set of proportions was then smoothed to give a sequence of adjusted population proportions that tended to preserve actual time trends in proportions while reducing the sampling variability of these estimates over time. The adjusted proportions were constrained by requiring that the sum of adjusted proportions across all subgroups for each year and reporting variable (formed by the categories of the variable and the smoothing cells) total one. For example, the sum of adjusted proportions for male and female 13-year-olds in 1972 had to equal one.

The sum of the adjusted proportions across the smoothing cells for a given year and reporting category provides an estimate of the proportion of eligibles in the population who were members of the reporting category. These sums are the proportions reported in Table B-2.

Once smoothed estimates of population proportions were obtained, respondent weights were adjusted so that adjusted performance estimates could be computed. As explained in Appendix A, the percentage of correct responses is estimated by dividing the sum of weights for students responding correctly to an exercise by the sum of weights for all students exposed to the exercise.

Exercise-level performance estimates are affected by both year-to-year sampling variability and within-year variability, because each exercise appears in only one package and is administered to a relatively small fraction of all

TABLE B-1. Definition of Smoothing Cells for Adjusting Population Proportions

Reporting Variable	Cell 1	Cell 2	Cell 3	Cell 4	Cell 5	Cell 6	Cell 7	Cell 8	Cell 9
Race	White	White	White	White	White	White	Black	Black	Other Race
Grade†	LTMG	LTMG	GEMG	GEMG	GEMG	GEMG	LTMG	GEMG	
Parental education††	PHS	NPHS	PHS	PHS	NPHS	NPHS			
Number of home reference items			4	<4	4	<4			

†*LTMG = Less than Modal Grade, GEMG = Modal Grade or Greater.*
††*PHS = Post High School; NPHS = High School or Less.*

TABLE B-2. Estimated Population Proportions of National Assessment Reporting Groups for Ages 9, 13 and 17 in 1969-70, 1972-73 and 1976-77

Reporting Groups	1969-70			1972-73			1976-77		
	Age 9	Age 13	Age 17	Age 9	Age 13	Age 17	Age 9	Age 13	Age 17
Sex									
Male	.495	.498	.489	.499	.500	.492	.502	.497	.490
Female	.505	.502	.511	.501	.500	.508	.498	.503	.510
Race									
White	.843	.851	.876	.808	.824	.853	.812	.808	.836
Black	.133	.132	.109	.141	.128	.112	.128	.135	.116
Other	.024	.017	.015	.051	.048	.035	.060	.057	.048
Region									
Northeast	.251	.245	.244	.260	.249	.268	.252	.257	.249
Southeast	.213	.223	.196	.224	.225	.198	.225	.224	.199
Central	.295	.291	.303	.275	.284	.292	.273	.273	.308
West	.241	.241	.258	.241	.242	.242	.250	.246	.243
Parental education									
Not graduated high school	.103	.154	.210	.095	.148	.173	.090	.134	.151
Graduated high school	.231	.314	.326	.220	.307	.317	.246	.328	.333
Post high school	.341	.412	.416	.325	.406	.463	.323	.408	.469
Unknown	.325	.121	.047	.360	.138	.047	.340	.129	.047
Type of community									
Extreme rural	.086	.096	.088	.085	.095	.081	.092	.103	.081
Low metro	.066	.088	.095	.077	.077	.096	.072	.071	.085
High metro	.124	.118	.140	.126	.118	.121	.102	.110	.102
Other	.724	.698	.677	.712	.710	.702	.734	.716	.732
Size of community									
Big city	.219	.218	.223	.209	.193	.183	.179	.173	.169
Fringes around big cities	.217	.207	.235	.224	.232	.252	.201	.185	.230
Medium city	.135	.144	.142	.139	.142	.143	.146	.132	.146
Smaller places	.428	.431	.399	.428	.433	.422	.474	.510	.455
Grade in school									
<3, <7, 10	.013	.033	.018	.010	.027	.017	.006	.021	.015
3, 7, 10	.232	.239	.125	.230	.246	.127	.232	.251	.136
4, 8, 11	.731	.715	.724	.747	.717	.728	.751	.720	.749
>4, >8, 12	.008	.011	.133	.006	.010	.127	.006	.008	.100
Other	.016	.002	.000	.007	.001	.001	.004	.000	.000

respondents. For example, in 1976 ten packages of exercises were administered to 13-year-olds. Smoothed population proportion estimates were based on 25,653 13-year-olds, but each exercise-specific performance estimate is based on the approximately 2,565 13-year-olds who took a particular package. Consequently, respondent weights were adjusted to dampen both between-year variability and package-to-package variability within an assessment year.

Respondent weights were adjusted separately for every reporting category by assessment year, age group and package combinations. To simplify the explanation, the adjustment process is described for male 13-year-olds who were administered package 1 (of 10) in 1976. The same process applies to all other combinations of reporting categories, ages, packages and assessments.

Weight sums were computed for the male 13-year-olds (who took package 1 in 1976) falling into each of the nine smoothing cells and converted to proportions by dividing by the sum of weights in all nine smoothing cells. An adjustment factor was then computed for each smoothing cell by dividing the smoothed proportion for that cell by the package proportion for the cell, as shown in Table B-3. The weight for each respondent (male 13-year-olds who took package 1 in 1976) in a smoothing cell was multiplied by the adjustment factor for the cell. Adjusted performance estimates were then computed with the adjusted weights.

The result of the smoothing and weight-adjustment process is that the estimated reporting-group proportions are identical for all packages (and exercises) in a particular age group and assessment year combination. More importantly, both adjusted performance estimates and changes in those estimates appear to be somewhat less susceptible to sampling variability, both across and within years. At the present time weighting class and other adjustment procedures continue to be evaluated to determine whether the increased precision in performance estimates is large enough to warrant the considerable additional costs involved.

TABLE B-3. Smoothing-Cell Proportions, Smoothed Proportions and Weight Adjustment Factors: Hypothetical Example for 13-Year-Old Males Taking Package 1 in 1976

	Smoothing Cells								
	Cell 1	Cell 2	Cell 3	Cell 4	Cell 5	Cell 6	Cell 7	Cell 8	Cell 9
Adjusted, smoothed proportions for male 13-year-olds in 1976	.090	.156	.208	.073	.158	.139	.047	.074	.054
Unadjusted proportions for male 13-year-olds in 1976 taking package 1†	.088	.154	.204	.072	.143	.136	.054	.096	.052
Weight adjustment factor (adjusted/unadjusted)	1.022	1.018	1.018	1.019	1.104	1.022	.860	.777	1.029

†*Unadjusted proportions are hypothetical because package-specific values were not printed. The values shown are for all 10 13-year-old packages administered in 1976. They are close to the proportions for any specific package.*

APPENDIX C

CHANGES IN PROCEDURES BETWEEN ASSESSMENTS

As with any sample survey, National Assessment results are subject to both sampling and nonsampling error. Sampling errors occur because responses are obtained only from a sample, not from the entire population. Nonsampling errors are unwanted variations in responses that might come from many sources in an assessment: the arrangement of exercises in packages, variability among exercise administrators, differing motivation levels of respondents, errors in recording responses and errors in data processing procedures, among others. When assessing change, we hold constant as many conditions as possible so that the nonsampling errors in the first assessment will cancel out those in the second assessment when the difference in achievement is computed.

However, it is not possible to control all sources of nonsampling error. Some conditions did change over the course of the three science assessments. This appendix describes changes in definitions of reporting variables, data collection procedures, "I don't know" responses and nonresponse to exercises. Comparative data on released versus unreleased exercises is also included.

Definition of Variables

Parental Education

The wording of the questions asking for level of parents' education was changed slightly after the first assessment. In 1969-70 respondents were asked, "How far did your father, or the man living in your home who acts as your father, go in school?" A similar question was asked about the respondent's mother. In subsequent assessments, the wording was simplified to: "How much school did your father complete?" with a similar question about the mother's schooling. Only the form of the question was changed; the response categories were not. After the 1972-73 assessment, results for changes in achievement by parental-education categories were not reported in the main body of change reports. However, because the results across the three assessments have been highly consistent despite the change in wording, they have been included in this report.

The proportion of respondents who did not report an education level for either parent has been high for 9-year-olds (about one-third) and lower for 13- and 17-year-olds (about 10 and 5%, respectively). Achievement of respondents in the unknown parental-education category is always lower than for any other category. Whether the low achievement of this group reflects lower

ability, lower parental interest or influence, motivational problems in the assessment situation or some other factor is not known.

Race

In 1969-70, exercise administrators visually identified respondents as white, black or other. In 1972-73 and 1976-77, partially in response to persons and organizations wanting information about other racial or ethnic groups, the exercise administrators were asked to classify respondents into one of five categories -- white, black, Puerto Rican, Mexican-American or other -- using visual identification and surname. When there was a question, the administrator was advised to determine the language or dialect the student spoke. In all cases, Puerto Rican and Mexican-American identification took priority over other categories.

The degree to which categorization into Puerto Rican and Mexican-American groups -- taking precedence over racial identification -- affected the racial categories themselves is not directly known. The proportions of whites and blacks have been within the range of sampling variability,[1] and group differences in achievement were quite consistent over the assessments. There may still be a small effect due to the change in definitions.

Community Size

In all three assessments community-size definitions were based on 1970 census data. Community characteristics have changed to some extent since 1970. Because of annexations, migration, births, etc., some smaller places have become more like medium cities or fringes of big cities, while some medium cities have become more like big cities or fringes of big cities. Data from Current Population Survey reports[2] indicate that the changes, while real, probably have not been large enough to seriously affect results for National Assessment categories. The 1980 census will provide more detailed data on community characteristics and migration between various geographic subpopulations. Analysis of population trends and their relationship to performance trends will be a major part of National Assessment's analysis and research effort in future assessments.

---

[1]See Appendix B for estimated population proportions.

[2]U.S. Bureau of the Census, "Mobility of the Population of the United States: March 1970 to March 1973," Current Population Reports, Series P-20, No. 262 (Washington, D.C.: U.S. Government Printing Office, 1974); U.S. Bureau of the Census, "Geographic Mobility: March 1975 to March 1977," Current Population Reports, Series P-20, No. 320 (Washington, D.C.: U.S. Government Printing Office, 1978).

Type of Community

In each assessment principals in sample schools were asked to estimate the proportion of adults in each of the following categories for the school attendance area:

A. Professional and managerial
B. Sales, clerical, technical and skilled
C. Factory and other blue collar
D. Farm workers
E. Not regularly employed
F. Welfare

Missing data were estimated from 1970 census reports.

Using these categories, rural, low-metro and high-metro indexes were then constructed for each school:

Rural: D-(C+2A)
Low metro: E+F-A
High metro: A-(C+D+E+F)

At each age, schools were excluded from the extreme-rural category if they were not in the smaller-places community-size category or if the principal reported that any students came from places of greater than 10,000 population. Remaining schools in this category that contained the 10% of the total sample highest on the extreme-rural index were classified as extreme rural. Only schools in big-city or fringes-around-big-cities categories were eligible for the high- and low-metro classification. Eligible schools containing the 10% of the sample highest on the high- or low-metro indexes were classified as high or low metro, respectively.

The extreme-type-of-community definitions have proved useful in identifying a constant percentage of respondents that are likely to be from opposite extremes on a rural-urban continuum and, within urban schools, at opposite extremes of a socioeconomic continuum. The populations represented each year are slightly different. The categories each year represent the most extreme 10% of students in that year's sample. If a particular year's sample happens to be less rural than previously defined, for example, then extreme rural will cover a less rural population that year. Also, the sample design used in each of the science assessments has defined and oversampled rural and low-socioeconomic areas somewhat differently. To the extent that National Assessment is more successful in oversampling these areas, 10% of the sample covers a smaller proportion of the extreme-rural and low-metro populations (and, conversely, a larger proportion of the high-metro population).

One other caution should be observed in interpreting extreme-type-of-community data. The older age groups' mean achievement is generally closer to the nation than is the younger age group's mean achievement. This phenomenon might be partly due to the larger size and heterogeneity of secondary-school attendance areas when compared to those of elementary schools.

Data Collection Procedures

Data Collection Staff

The first assessment of science occurred during the first year that National Assessment collected data. The second and third assessments of science took place in the fourth and eighth years of data collection, respectively, by which time several improvements in field operations had been made. For example, the field staff used in later assessments had more experience and better training than did the staff in the initial assessment year. Better quality-control procedures were also implemented so that the field staff could be contacted quickly and instructed about procedural changes if there were difficulties in administration.

Learning-Area Mix

In 1969-70, science exercises were administered in packages also containing writing and citizenship exercises; in 1972-73, science was administered with mathematics. Most of the citizenship and writing exercises were short-answer or essay exercises, while most of the mathematics exercises required respondents to compute and record their own answers. Most of the mathematics exercises were short; the citizenship and writing exercises were longer. Although the total testing time (about 40 minutes per respondent) was the same in each assessment, responding to many short exercises rather than a few long ones may have had an effect on performance. The 1976-77 packages contained only science exercises, most of which were multiple-choice. For the first time in a science assessment, most packages contained experience or attitude inventories.

Taping

There were some slight variations in the taping of the exercises. New tapes were made in each assessment because different combinations of science exercises and learning areas were assessed each time. A different announcer was used in the first assessment than in the last two, but in each assessment the announcer read clearly and at a constant rate. Tapescripts for change exercises were kept as constant as possible (including errors), but there were slight changes in the introductory remarks, transitional remarks between exercises and instructions on the use of the "I don't know" response.

All 1972-73 taping conventions were replicated as closely as possible on exercises for the 1976-77 assessment. In the second assessment, the announcer said, at the end of each exercise, "If you do not know the answer, please fill in the oval beside 'I don't know.'" Slightly different conventions were used for new exercises in 1976-77. "I don't know" was read immediately after the other response choices at age 9; it was not read at all at ages 13 and 17. To minimize the effect of these changes, old exercises were clustered at either the beginning or end of each package. They were not segregated in separate packages because of the increased precision of summaries when exercises are spread over multiple packages.

Printing

Instructions on the bottom of each page telling respondents to stop or to continue on to the next page were given added visual emphasis in the second assessment: "Stop" appeared in an octagon and "Please Continue on the Next Page" appeared in an arrow. In addition, there were slight changes in the sizes of type faces used in the two assessments. Both sets of printing have been judged by National Assessment's reading consultants to be easily readable at the appropriate age levels. Printing was essentially identical in the second and third assessments.

Mode of Administration

Most of the assessments conducted by National Assessment have contained both individual (one-to-one interviews) and group administrations. All exercises used to measure changes in achievement between the first two assessments were group administered. In 1976-77 all exercise packages were group administered. However, two exercises used to measure change between 1973 and 1977 were administered individually to 9-year-olds in 1973. Exercise 202029 asked whether water would weigh more, the same or less when frozen. Respondents were then asked to explain their choice of more, the same, or less. Only the multiple-choice portion was used in change summaries. Exercise 202072 was a multiple-choice exercise that required students to pick the picture of a can that might contain botulism poison. Changes in the percentage of correct responses to both exercises were negative (-8.5% and -3.1%, respectively), while the average change for all exercises was essentially zero. However, neither change figure appeared to be unreasonably large when compared with changes for the other exercises.

In the first two assessments, group administrations were limited to 12 students. In 1976-77, the planned average group size was set at 16, with a range of 10 to 35 students. Some problems with overcrowding were encountered in the larger sessions.

Position in Package

In both the second and third science assessments, science exercises were reassigned to assessment packages. In 1972-73, new and old science exercises were mixed with mathematics exercises. In packaging exercises, National Assessment staff attempted to balance difficulty level, objective, content type and other variables across packages with the constraints of fixed total assessment time for each package of exercises, and no exercise in a package could provide the answer to any other exercise. In preparation for the 1976-77 assessment, nearly all change exercises from a 1972-73 package were put together in either the beginning or end of a 1976-77 package. There were some exceptions due to differing numbers of packages between assessments and other constraints.

Having all change exercises at the beginning or end of a package represented a major departure from prior assessment practice. If there were biases

associated with package location, the validity of change measures would be jeopardized. It has been suggested that examinees might do poorly on the first exercise (or exercises) in a testing situation because of the initial tension examinees sometimes experience.[3] In addition, performance on the last exercise (or exercises) might be lower than expected if some examinees do not have time to complete them.

National Assessment attempts to control the effects of exercise position in a package by presenting an audio explanation of what the assessment is, how results will be used, and one or more example exercises before actual assessment begins. Further, exercises are presented on audiotapes to pace respondents through to the end of the packages.

After the second science assessment, National Assessment staff analyzed results for exercises that were first or last in a package in either assessment. There was little relationship between position and changes in achievement.[4] A small controlled experiment on position and format was included in the 1973-74 assessment of writing and career and occupational development. Even with approximately 7,500 respondents per treatment condition, no systematic position effects were detected.[5]

Position-in-package effects between the second and third assessments were investigated by dividing packages into thirds and classifying exercises by location in the second and third assessments. Exercise administrations in the last part of a package are sometimes lost when sessions start late, schools close early, etc., so means for both correct response and nonresponse were computed by package location. Numbers of exercises, mean changes in percentages of correct responses and standard deviations[6] are listed in Table C-1 and plotted in Figure C-1. The same statistics for nonresponse[7] are listed in Table C-2

---

[3]R.L. Ebel, Measuring Educational Achievement (Englewood Cliffs, N.J.: Prentice Hall, 1965).

[4]See Science Technical Report: Summary Volume, Report 04-S-21, pp. 100-104 for detailed documentation on the position in package analysis (available from National Assessment offices).

[5]N. Burton et al., "The Effect of Position and Format on the Difficulty of Assessment Exercises," paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, April 1976.

[6]Standard deviations are included as indicators of the variation in exercise-level changes in each cell. Because not all exercises were administered to all students, they are not valid statistics for testing differences between cells.

[7]The nonresponse reported in the appendix is nonresponse to exercises for respondents who were present for package administrations. Failure to participate in the assessment because of school or student refusal is treated in Appendix D.

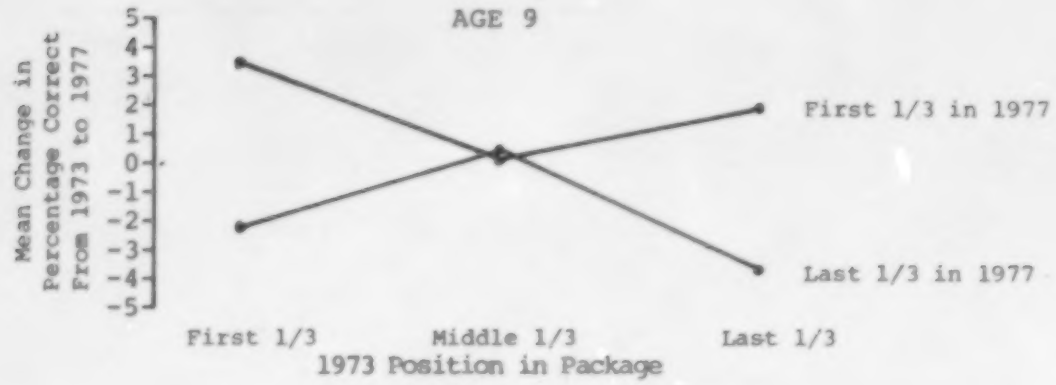
TABLE C-1. Mean Change in Percentages of Correct Responses Between 1972-73 and 1976-77 by Package Location in 1972-73 and 1976-77, Ages 9, 13 and 17

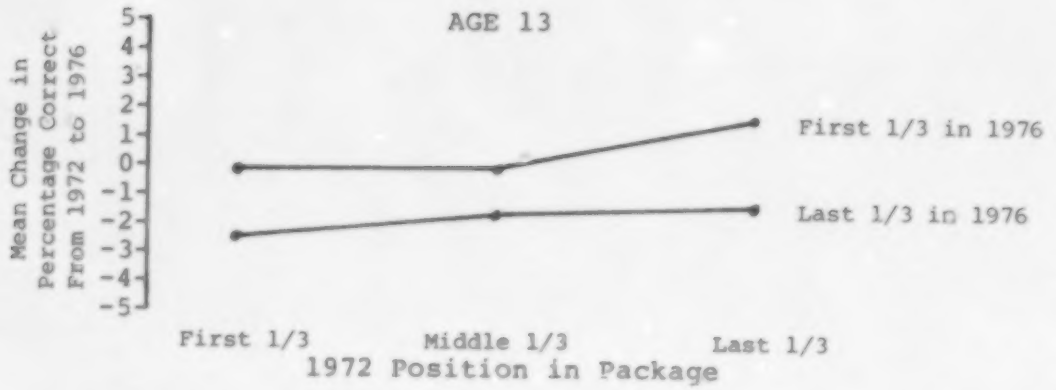
Age	1976-77 Package Location		1972-73 Package Location: First 1/3	Middle 1/3	Last 1/3	Total
9	First 1/3	Mean change, (S.D.)†	3.5 (3.0)†	.1 (3.2)	1.8 (4.3)	1.8 (3.7)
		Number of exercises	13	13	13	39
	Last 1/3	Mean change, (S.D.)	-2.2 (5.2)	.4 (4.9)	-3.6 (3.5)	-1.9 (4.7)
		Number of exercises	9	11	12	32
	Total	Mean change, (S.D.)	1.1 (4.9)	.2 (4.0)	- .8 (4.7)	.1 (4.5)
		Number of exercises	22	24	25	71
13	First 1/3	Mean change, (S.D.)	- .1 (3.3)	- .2 (4.0)	1.5 (2.7)	.3 (3.4)
		Number of exercises	11	14	11	36
	Last 1/3	Mean change, (S.D.)	-2.4 (5.4)	-1.7 (5.4)	-1.5 (6.5)	-1.8 (5.8)
		Number of exercises	12	10	17	39
	Total	Mean change, (S.D.)	-1.3 (4.6)	- .9 (4.6)	- .3 (5.5)	- .8 (4.9)
		Number of exercises	23	24	28	75
17	First 1/3	Mean change, (S.D.)	-1.3 (3.0)	-1.9 (3.6)	-1.1 (3.5)	-1.4 (3.3)
		Number of exercises	15	15	13	43
	Last 1/3	Mean change, (S.D.)	-2.2 (3.2)	- .9 (3.5)	-3.0 (3.0)	-2.2 (3.2)
		Number of exercises	9	7	11	27
	Total	Mean change, (S.D.)	-1.7 (3.0)	-1.6 (3.6)	-1.9 (3.4)	-1.7 (3.3)
		Number of exercises	24	22	24	70

†*S.D. = Standard Deviation.*

Note: Data in this table were computed prior to weight smoothing (see Appendix B). Consequently, mean changes differ slightly from those reported in Chapter 1.

FIGURE C-1. Mean Change in Percentage of Correct Response From 1972-73 to 1976-77 by Position in Package, Ages 9, 13 and 17





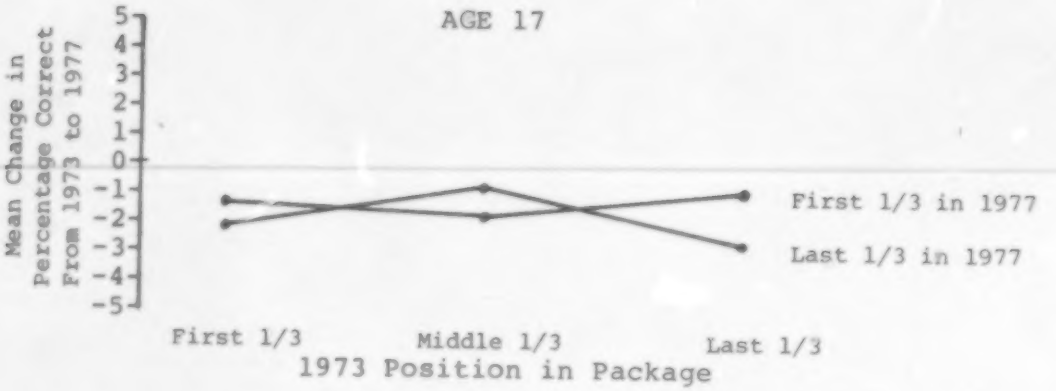


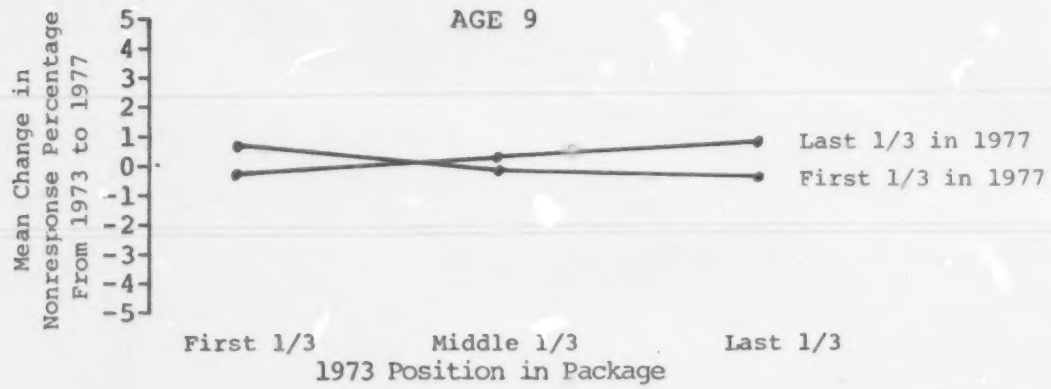
TABLE C-2. Mean Change in Percentages of Nonresponses Between 1972-73 and 1976-77 by Package Location in 1972-73 and 1976-77, Ages 9, 13 and 17

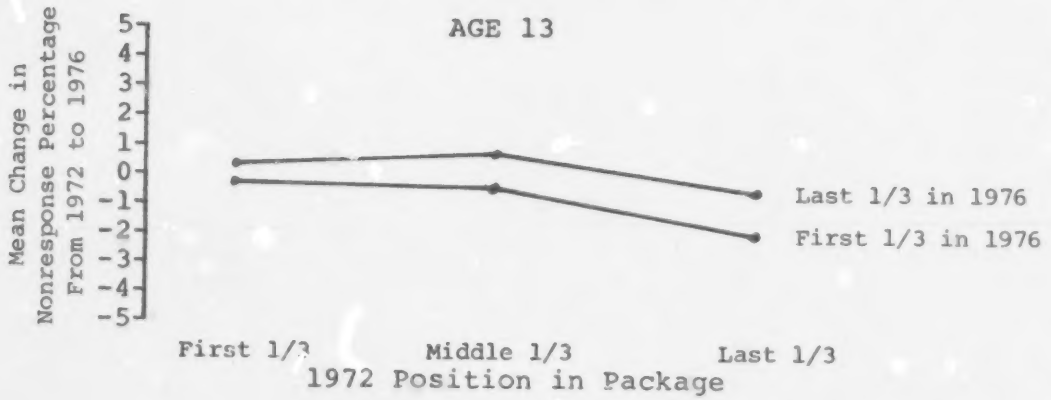
Age	1976-77 Package Location		1972-73 Package Location: First 1/3	Middle 1/3	Last 1/3	Total
9	First 1/3	Mean change, (S.D.)†	.7 (3.3)†	- .1 (.2)	- .3 (.8)	.0 (2.0)
		Number of exercises	12	13	13	38
	Last 1/3	Mean change, (S.D.)	- .3 (.6)	.3 (1.4)	.8 (.3)	.3 (1.0)
		Number of exercises	9	10	12	31
	Total	Mean change, (S.D.)	.3 (2.6)	.1 (1.0)	.2 (.8)	.1 (1.6)
		Number of exercises	21	23	25	69
13	First 1/3	Mean change, (S.D.)	- .2 (.8)	- .3 (.4)	-1.8 (1.3)	- .8 (1.1)
		Number of exercises	11	13	11	35
	Last 1/3	Mean change, (S.D.)	.2 (.7)	.8 (2.1)	- .1 (2.4)	.2 (1.9)
		Number of exercises	11	10	17	38
	Total	Mean change, (S.D.)	.0 (.8)	.2 (1.5)	- .8 (2.2)	- .2 (1.7)
		Number of exercises	22	23	28	73
17	First 1/3	Mean change, (S.D.)	.4 (.5)	- .9 (3.5)	- .6 (.4)	- .3 (2.0)
		Number of exercises	15	13	13	41
	Last 1/3	Mean change, (S.D.)	1.2 (.5)	1.6 (.7)	1.9 (.7)	1.6 (.7)
		Number of exercises	9	7	11	27
	Total	Mean change, (S.D.)	.7 (.6)	.0 (3.0)	.5 (1.4)	.4 (1.9)
		Number of exercises	24	20	28	68

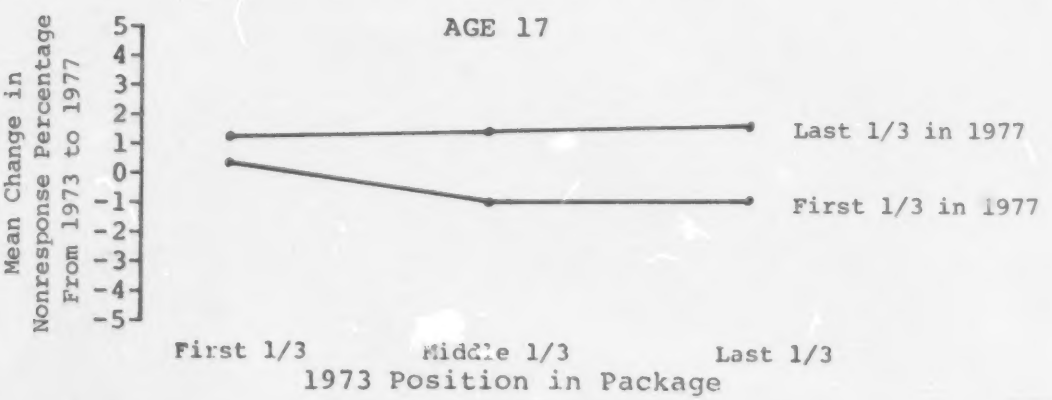
†S.D. = Standard Deviation.

Note: Data in this table were computed prior to weight smoothing (see Appendix B). Consequently, mean changes differ slightly from those computed from smoothed weights.

FIGURE C-2. Mean Change in Percentage of Nonresponse From 1972-73 to 1976-77 by Position in Package, Ages 9, 13 and 17







and plotted in Figure C-2. We expected that exercises either at the beginning or at the end of a package in both assessments would provide the best control of nonsampling error. Results were not always consistent with this expectation.

At age 9, exercises appearing in the first third of a package in both assessments had a positive change of 3½ percentage points, while those appearing in the last third of a package in both assessments had a slightly larger negative change. At age 13, mean changes for exercises appearing in the first third of 1976-77 packages were less negative than for exercises appearing in the last third of 1976-77 exercises. At age 17, differences between means were small and inconsistent.

The relationship between changes in percentages of correct responses and position in package was not sufficiently clear to merit further action. The numbers of exercises in various positions were well balanced across the second and third assessments. Exercises were not randomly assigned to location in either assessment, and the differences observed might have been confounded by content or some other unknown variable.

Changes in nonresponse were slightly but consistently higher for exercises appearing in the last third of 1976-77 exercises. The only exception occurred at age 9, where a large increase in nonresponse to one open-ended exercise caused a reversal in mean changes for exercises in the first third of 1972-73. Deleting that one exercise makes the two means identical. After inspecting mean changes in nonresponse (Table C-2 and Figure C-2) and exercise-by-exercise plots of changes in position in package, nonresponse was dropped from further consideration. The rates of nonresponse were too small and too unrelated to package location to merit any adjustment of correct response change statistics.

"I Don't Know" Responses and Nonresponse

National Assessment emphasizes to respondents that it is not a test in the usual sense and scores are not reported for individuals. Exercises are presented on audiotapes to help ensure exposure to all exercises, and the response choice "I don't know" is included among the possible choices on all cognitive multiple-choice exercises to minimize guessing.[8]

Table C-3 contains the mean percentages of "I don't know" responses in 1969-70 and 1972-73 for exercises used to measure changes in achievement between the first two assessments, and similar data for exercises used to measure changes in achievement between the second and third assessments.

---

[8]N. Burton et al., "The Effect of Position and Format on the Difficulty of Assessment Exercises."

TABLE C-3. Mean Percentage Responding "I Don't Know" in 1969-70, 1972-73 and 1976-77 for Exercises Used to Measure Change From 1969-70 to 1972-73 and From 1972-73 to 1976-77, Ages 9, 13 and 17

	Age 9 Percentages				Age 13 Percentages				Age 17 Percentages			
	92 Exercises		66 Exercises		67 Exercises		69 Exercises		63 Exercises		64 Exercises	
	1970	1973	1973	1977	1969	1972	1972	1976	1969	1973	1973	1977
Nation	6.2	6.0	7.6	10.3	6.9	8.1	7.4	8.7	12.9	14.0	9.9	11.4
Region												
Northeast	6.0	5.4	7.1	8.9	6.5	7.8	6.9	8.0	12.1	13.5	9.7	10.3
Southeast	6.2	6.8	8.4	11.9	6.1	7.3	7.2	8.0	12.0	13.3	9.6	11.8
Central	6.1	5.6	6.9	10.6	6.9	8.5	8.1	9.1	13.7	14.3	9.9	11.6
West	6.5	6.1	7.9	9.9	8.1	8.5	7.6	9.8	13.2	14.6	10.4	12.1
Sex												
Male	5.2	5.1	6.5	9.2	6.1	7.1	6.5	7.6	11.3	12.2	8.6	9.9
Female	7.2	6.8	8.7	11.5	7.7	9.0	8.4	9.8	14.4	15.6	11.3	13.1
Race												
White	5.9	5.6	7.1	10.0	6.6	7.6	7.1	8.4	12.9	13.6	9.5	11.1
Black	7.4	7.1	8.9	11.8	7.9	9.9	9.3	9.8	12.3	15.3	11.1	13.6
Type of community												
Extreme rural	6.7	6.5	7.7	10.6	8.3	7.5	7.2	9.7	12.6	14.4	10.0	12.4
Low metro	7.7	7.6	9.6	14.0	8.8	10.4	9.4	9.0	12.8	15.5	11.1	12.1
High metro	4.4	4.7	6.4	7.5	6.6	8.0	7.2	8.1	12.6	13.2	9.2	10.1

There was a slight but fairly consistent increase in usage across the three assessments. There was an increase in "I don't know" responses with age on exercises used in both 1969-70 and 1972-73; that trend is not apparent on exercises used in both 1972-73 and 1976-77. Reporting-group usage of the "I don't know" response mirrored achievement trends fairly closely. Sex, race and community-type differences were all the opposite of achievement differences, while the pattern for regional groups is not clear.

Table C-4 shows the mean percentages of nonresponse in 1969-70 and in 1972-73 on exercises used to measure changes in achievement between the first two assessments, and similar data for exercises used to measure changes in achievement from 1972-73 to 1976-77. The mean percentage of exercise nonresponse ranged from approximately ½ to 1% across all ages and assessments. Nonresponse for various reporting groups tends to mirror achievement patterns. For example, blacks and low-metro students have somewhat higher nonresponse rates than whites and high-metro students, just the opposite of the achievement results. The trend is less clear for regional and sex groups, where achievement differences were smaller than for race and type of community.

Released and Reassessed Exercises

Most National Assessment change measures are based on exercises that have never been released for public use. The 1969-70 to 1972-73 change summaries contained 10, 6 and 4 previously released exercises at ages 9, 13 and 17, respectively. Analyses of changes in achievement on released versus unreleased exercises were inconclusive.[9] At ages 9 and 17, achievement on released exercises declined at the same rate as that on unreleased exercises, while at age 13, the changes in achievement were generally positive for both types of exercises. However, one of the exercises showing a large, positive change at age 13 exhibited a large, negative change at age 17.

A number of previously released exercises were included in the 1976-77 science assessment; however, none has been included in change summaries. Almost all of those previously released exercises were released after the 1969-70 assessment. Change results for those exercises and unreleased exercises from the 1969-70 assessment are shown in Tables C-5 to C-7.[10] Because differential changes have been observed in biology and physical science exercises, results are given by type of science as well as for all exercises.

For all exercises, changes in the percentages of correct responses between 1969-70 and 1976-77 were quite similar for released and unreleased exercises.

---

[9]See Science Technical Report: Summary Volume, Report 04-S-21, pp. 105-108 for additional details.

[10]All results were computed prior to weight smoothing; change statistics for unreleased exercises differ slightly from those reported in Chapter 2.

TABLE C-4. Mean Percentage of Nonresponse in 1969-70, 1972-73 and 1976-77 for Exercises Used to Measure Change From 1969-70 to 1972-73 and From 1972-73 to 1976-77, Ages 9, 13 and 17

	Age 9 Percentages				Age 13 Percentages				Age 17 Percentages			
	92		66		67		68		64		62	
	Exercises		Exercises		Exercises		Exercises		Exercises		Exercises	
	1970	1973	1973	1977	1969	1972	1972	1976	1969	1973	1973	1977
Nation	.4	.5	.6	.8	.2	.8	1.1	.9	.3	1.1	.8	1.1
Region												
Northeast	.3	.6	.5	.7	.3	.3	.4	.6	.2	.8	.8	.9
Southeast	.4	.6	.8	1.0	.2	.3	.8	.5	.2	.5	.5	.8
Central	.3	.4	.6	.8	.1	.7	.9	1.4	.1	.5	.5	1.3
West	.4	.5	.7	.9	.1	1.7	2.4	1.2	.3	2.5	1.3	1.3
Sex												
Male	.5	.6	.8	.9	.2	.9	1.2	.9	.2	1.2	.9	1.1
Female	.4	.5	.6	.8	.2	.7	1.0	.9	.3	1.1	.7	1.1
Race												
White	.2	.4	.5	.6	.1	.7	1.0	.6	.2	.9	.7	.9
Black	1.6	1.1	1.6	2.2	.4	.9	1.6	2.4	.5	1.6	1.4	2.3
Type of community												
Extreme rural	.1	.4	.6	.6	.3	1.4	1.7	.3	.2	.1	.1	.6
Low metro	1.5	1.6	1.9	1.7	.4	2.0	3.0	4.4	.4	2.1	1.6	3.5
High metro	.2	.3	.4	.6	.1	.1	.4	.9	.2	1.6	1.2	.5

TABLE C-5. Mean Changes in Percentages of Correct Responses From 1970 to 1977 for Released and Unreleased Exercises by Content Classification, Age 9

Classification		Number of Exercises	Change 1970 to 1973	Change 1973 to 1977	Total Change
Biology	Released after 1970 assessment	5	- .3	-1.4	-1.7
	Standard error		(1.0)	(.9)	(1.0)
	Unreleased	11	-1.1	.8	- .3
	Standard error		(.7)	(.8)	(.8)
Physical science	Released after 1970 assessment	8	-2.3	-0.1	-2.4
	Standard error		(1.0)	(1.1)	(1.0)
	Unreleased	16	-2.4	-2.3	-4.7
	Standard error		(.7)	(.7)	(.7)
Unclassified	Released after 1970 assessment	0	--	--	--
	Standard error		--	--	--
	Unreleased	3	1.6	3.1	4.8
	Standard error		(1.0)	(1.1)	(1.1)
Total	Released after 1970 assessment	13	-1.5	- .6	-2.2
	Standard error		(.8)	(.9)	(.9)
	Unreleased	30	-1.5	- .6	-2.1
	Standard error		(.6)	(.6)	(.6)

TABLE C-6. Mean Changes in Percentages of Correct Responses From 1969 to 1976 for Released and Unreleased Exercises by Content Classification, Age 13

Classification		Number of Exercises	Change 1969 to 1972	Change 1972 to 1976	Total Change
Biology	Released after 1969 assessment	4	-2.0	-1.8	-3.8
	Standard error		(1.3)	(1.4)	(1.4)
	Unreleased	7	- .7	2.2	1.5
	Standard error		(.8)	(.8)	(.9)
Physical science	Released after 1969 assessment	9	-1.6	-2.1	-3.7
	Standard error		(.9)	(.9)	(.9)
	Unreleased	13	-3.3	-2.1	-5.4
	Standard error		(.8)	(.8)	(.7)
Unclassified	Released after 1969 assessment	1	-4.1	-3.6	-7.7
	Standard error		(2.4)	(2.4)	(2.3)
	Unreleased	3	.4	-9.5	-9.0
	Standard error		(1.2)	(1.4)	(1.4)
Total	Released after 1969 assessment	14	-1.9	-2.1	-4.0
	Standard error		(.9)	(1.0)	(.9)
	Unreleased	23	-2.0	-1.8	-3.8
	Standard error		(.6)	(.7)	(.7)

TABLE C-7. Mean Changes in Percentages of Correct Responses From 1969 to 1977 for Released and Unreleased Exercises by Content Classification, Age 17

Classification		Number of Exercises	Change 1969 to 1973	Change 1973 to 1977	Total Change
Biology	Released after 1969 assessment	1	-3.8	-8.2	-12.0
	Standard error		(2.4)	(2.0)	(2.4)
	Unreleased	8	-2.6	- .6	-3.3
	Standard error		(.7)	(.7)	(.7)
Physical science	Released after 1969 assessment	5	-3.0	-3.3	-6.3
	Standard error		(1.0)	(1.0)	(1.1)
	Unreleased	13	-2.7	-3.6	-6.3
	Standard error		(.8)	(.8)	(.8)
Unclassified	Released after 1969 assessment	2	-5.4	-1.7	-7.1
	Standard error		(1.6)	(1.5)	(1.6)
	Unreleased	2	-4.2	- .7	-4.8
	Standard error		(1.2)	(1.0)	(1.2)
Total	Released after 1969 assessment	8	-3.7	-3.5	-7.2
	Standard error		(.9)	(.9)	(1.0)
	Unreleased	23	-2.8	-2.3	-5.2
	Standard error		(.6)	(.6)	(.6)

When classified by type of science, results are less consistent. At all three ages, released biology exercises showed larger declines than did unreleased biology exercises. In physical science, declines were somewhat lower for released than for unreleased exercises. There were very few unclassified exercises, and changes on these were quite similar for released and unreleased exercises at ages 13 and 17. Thus, for the small number of released science exercises on which National Assessment has repeated change measures, release for public use may not have had much effect on the percentage of correct responses.

National Assessment has no control over the use of released exercises. At any time, a specific exercise may be used in other assessment or testing programs or reproduced in newspapers, journal articles or textbooks. If large numbers of students are exposed to the exact content of the exercise, it is irreparably contaminated for measuring changes in achievement. Consequently, National Assessment's reuse of previously released exercises is minimal.

APPENDIX D

NONRESPONSE IN ASSESSMENT SAMPLES

In addition to sampling variability, estimates of population values computed from sample surveys might be subject to random error and systematic bias. Systematic bias, or nonrandom error, might result from estimation procedures, errors inherent in measurement and data collection procedures, and nonresponse. Sampling variability and random error are discussed in Chapter 1 and nonrandom errors are described in Appendix C. This appendix examines nonresponse in the 1969-70, 1972-73 and 1976-77 assessments. Since nonresponse rates at ages 9 and 13 are relatively small, the following discussion concerns 17-year-olds' response rates only.

Bias due to nonresponse is present in virtually every sample survey but is frequently ignored since it is difficult to estimate its size. A variety of factors contribute to nonresponse. Nonrespondents might either be difficult to notify or reluctant to participate once they are notified; some might be absent from school during the entire contact period with item administrators. However, these nonrespondents can be important, since, if they respond differently than did the people actually included in the sample, estimates of percentage based solely on the sample are biased and not properly representative of the age population being assessed.

To provide some information about the size of the bias due to nonresponse in National Assessment surveys, the Research Triangle Institute, Raleigh, North Carolina, was asked to conduct a special study of nonrespondents during the 1972-73 assessment of science and mathematics. The study was conducted on the age population of eligible 17-year-olds who, at the time of the assessment, were listed as enrolled in school. Some of these students, in fact, were no longer attending school at the time of the assessment. Eligibles had to be English-speaking, physically and emotionally able to respond to exercises as administered and not residing in an institution.

The results of the nonresponse study[1] indicate that 17-year-olds listed as enrolled in schools but not appearing at the designated time of assessment can be divided into two different groups. The first group of nonrespondents,

---

[1]W.D. Kalsbeek et al., No Show Analysis, Final Report (Raleigh, N.C.: Research Triangle Institute, 1975); W.T. Rogers et al., "Assessment of Nonresponse Bias in Sample Surveys: An Example From National Assessment," Journal of Educational Measurement, Vol. 14, No. 4, 1977.

which comprises about 80% of the total nonrespondent group, did not appear for the assessment because of conflicting school activities or illness. The performance of this group was not very different from the performance of students assessed during the regularly scheduled sessions. The second group of nonrespondents, which comprises about 20% of the nonrespondents, do not appear to be available in the schools at any time. They attend infrequently if they attend at all (for practical purposes they have dropped out of school), or they have moved out of the school attendance area. In either case, these students should probably not have been listed in the in-school population of eligibles. This group, in contrast to the group of nonrespondents who were in fact attending school, performed more poorly on assessment questions than students assessed during the scheduled sessions.

The weights used by National Assessment to estimate the percentage of acceptable responses are adjusted for nonresponse. The adjustment assumes that the nonrespondents would perform, on the average, in a manner similar to those who did respond. However, the nonresponse study showed that the second group of nonrespondents, those enrolled in but not actually attending school, typically performed at a lower level than either those who did respond or the first group of nonrespondents. If the second group is included in the population of eligibles, the nonresponse adjustment procedure used by National Assessment would result in overestimates of the true percentages of acceptable responses.

Because the second group of students is effectively no longer attending school, it does not seem appropriate to include them in estimates for 17-year-olds in school. Thus, these students are not considered part of the population of eligibles and are excluded from the computations of percentage of the sample covered for 17-year-olds shown in Table D-1.

Including the second group of students and then reducing bias due to their nonresponse would require the location and testing of some of these individuals. The difficulty and costs associated with supplementary data collection of the nonrespondents not actually attending school are so great that this has not been a feasible alternative in recent years.

National Assessment continually evaluates its field procedures and has introduced new methods to lessen the effects of nonresponse. In the second and third assessments of science for 17-year-olds, item administrators used the day following a regularly scheduled assessment session to locate and assess nonrespondents. This helped to reduce the bias due to nonresponse of students enrolled in and attending school.

However, systematic bias in change measures can be introduced if the use of new procedures results in very different samples in different assessment years. Thus, measures of change from previous years are still based upon samples obtained using the old procedures. Measures intended for use in determining future changes are based on samples obtained using the new procedures.

Table D-1 shows the average sample coverage per package (booklet) of exercises administered in 1969-70, 1972-73 and 1976-77. The rate of coverage is

TABLE D-1. Number of Students Assessed and Percent of Sample Covered by Age, Assessment Year and Type of Administration

Year	Age	Type of Administration†	Number of Packages	Total Number of Students Assessed	Average Number Assessed Per Package	Average Sample Coverage in Percent
1969-70	9	G	8	19,468	2,434	88.0††
		I	2	3,713	1,856	89.1††
	13	G	9	21,696	2,411	85.6††
		I	3	5,568	1,856	87.2††
	17	G	11	22,913	2,083	74.5
		I	2	3,328	1,664	71.2
1972-73	9	G	7	18,638	2,663	91.0
		I	3	6,766	2,255	89.3
	13	G	9	23,507	2,612	84.6
		I	3	6,744	2,248	85.5
	17	G	11	25,865	2,351	73.6
		I	3	6,500	2,167	77.2
1976-77	9	G	7	17,345	2,478	88.6
	13	G	10	25,653	2,565	86.2
	17	G w/o F**	11	29,140	2,649	73.1
		G w/F**	11	34,514	3,137	83.7

†G indicates group-administered packages; I indicates package was administered to one student at a time.
††Complete records at ages 9 and 13 were not available. These are best estimates of sample coverage.
***W/F indicates group administration with follow up of nonrespondents attending school. W/O F indicates group administration without follow up of nonrespondents attending school.*

Note: Computations for 17-year-olds do not include those enrolled in school but not actually attending.

based on an estimated total eligible age population of students who are available in school -- for 17-year-olds, those enrolled minus the 20% estimated to be enrolled but unavailable in school. For completeness, figures are also shown for individual interview packages. Only one individually administered package at age 9 in 1972-73 contained exercises used to measure changes in achievement.

Figures for 17-year-olds include both a sample of 17-year-olds assessed according to old procedures (no attempts were made to contact and assess nonrespondents the following day) and a sample assessed according to the new procedures (attempts were made to contact and assess nonrespondents). Since the 1969-70 and 1972-73 samples did not include follow-up attempts, changes in percentages between assessments are based upon the 1976-77 sample that does not include the follow-up attempts. Changes toward future years will be based upon the sample that does include follow-up attempts.

☆ U. S. GOVERNMENT PRINTING OFFICE 1979 - 679-244/381 Reg. 8

NATIONAL ASSESSMENT OF EDUCATIONAL PROGRESS
Education Commission of the States

Dixy Lee Ray, Governor of Washington, Chairperson, Education Commission of the States
Warren G. Hill, Executive Director, Education Commission of the States
Roy H. Forbes, Director, National Assessment

All National Assessment reports and publications are available through NAEP offices at the address shown at the bottom. Some of the more recent results reports are also available at the Superintendent of Documents (SOD), usually at lower prices. To order from the SOD, write to Supt. of Documents, U.S. Government Printing Office, Washington, D.C. 20402. Check must accompany order. Allow four to eight weeks for delivery.

Reports ordered from National Assessment should be delivered within 12 days. Reports related to this report and available from National Assessment include:

BASIC LIFE SKILLS (special probe)

1st Assessment (1976-77)

08-BLS-21	*Basic Life Skills Technical Report*, April 1979	$ 9.80

ADULTS (special probe)

1st Assessment (1976-77)

08-H-01	*Checkup: A National Assessment of Health Awareness Among 17-Year-Olds and Young Adults*, September 1978	3.75
08-E-01	*Energy: Knowledge and Attitudes, A National Assessment of Energy Awareness Among Young Adults*, December 1978	3.75
----	*Technical Information and Data From the 1977 Young Adult Assessment of Health, Energy and Reading*, March 1979	15.00

SCIENCE

1st Assessment (1969-70)

Report 1	*Science: National Results*, July 1970	2.15
Report 4	*Science.* Results by sex, region and size of community, April 1971	1.00
Report 7	*Science.* Results by race, parental education, size and type of community; also balanced results for all groups, May 1973	2.70

2nd Assessment (1972-73)

04-S-01	*Selected Results From the National Assessments of Science: Energy Questions*, May 1975	1.45
04-S-02	*Selected Results From the National Assessments of Science: Scientific Principles and Procedures*, August 1975	3.00
04-S-03	*Selected Results From the National Assessments of Science: Attitude Questions*, October 1975	3.45
04-S-00	*National Assessments of Science, 1969 and 1973: A Capsule Description of Changes in Science Achievement*, February 1975	1.00
04-S-20	*Changes in Science Performance, 1969–73: Exercise Volume*, December 1975	25.00
04-S-20	*Changes in Science Performance, 1969–73: Exercise Volume, Appendix* (2 vols.), April 1977	25.00
04-S-21	*Science Technical Report: Summary Volume*, May 1977	16.80
BRS-1	*Science Achievement: Racial and Regional Trends, 1969-73*, March 1976	3.95

3rd Assessment (1976-77)

08-S-00	*Three National Assessments of Science: Changes in Achievement, 1969–77*, June 1978	2.40
08-S-01	*Science Achievement in the Schools: A Summary of Results From the 1976–77 National Assessment of Science*, December 1978	3.60
08-S-21	*Three Assessments of Science, 1969–77: Technical Summary*, April 1979	10.60
----	*The Third Assessment of Science, 1976–77: Released Exercise Set*, May 1978	18.90
----	*Technical Appendix to the Third Assessment of Science, 1976–77: Released Exercise Set*, December 1978	2.75

BACKGROUND REPORTS

BR-2	*Hispanic Student Achievement in Five Learning Areas: 1971–75.* Data for 9-, 13- and 17-year-olds in reading, mathematics, science, social studies and career and occupational development, May 1977	4.45
03/04-GIY	*General Information Yearbook.* A condensed description of the Assessment's methodology, December 1974	2.50

In addition to the above reports, National Assessment has produced reports in the areas of social studies, citizenship, writing, literature, reading, music, art and career and occupational development. A complete publications list and ordering information are available from the address below.

NATIONAL ASSESSMENT OF EDUCATIONAL PROGRESS
Suite 700, 1860 Lincoln Street
Denver, Colorado 80295